CTPEA

Cours : Statistique Mathématique

Professeur : Jean-Baptiste ANTENORD — Jeudi : 10h-12h

Année Académique : 2023-2024 — Vendredi : 9h30-10h30

Semestre : II

Niveau : Deuxième Année

Pré requis : Statistique Descriptive, Probabilités II

==========

Chapitre II : Test d'hypothèses statistiques

2.1 Problématique d'un test statistique et définitions.

2.1.1 Problématique d'un test

Soit un modèle paramétrique $(D_X, \boldsymbol{a}, P_\theta)^n$, $\theta \in \Theta \in IR^d$ avec $d \in IN^*$ où le paramètre θ est inconnu. Le statisticien ne cherche pas directement à inférer la valeur θ mais plutôt de savoir si θ appartient à un ensemble de paramètres $\Theta_0 \subseteq \Theta$: l'objectif d'un test est de décider si $\theta \in \Theta_0$, ou pas.

Exemple. Une des premières applications de la théorie des tests était liée au problème militaire de détection de la présence d'un missile à l'aide d'un radar. L'écho d'un radar est "grand" si un missile est présent et il est "petit" dans le cas contraire. Supposons qu'on observe un échantillon $(X_1, X_2, ..., X_n)$ d'échos de radar aux instants successifs *1,2, . . . , n.* Le caractère aléatoire de ces échos est lié aux effets de bruit de propagation d'ondes, des erreurs de mesure, etc... On se place dans le cadre d'un modèle paramétrique où $(X_1, X_2, ..., X_n)$ est issu d'un modèle P_θ avec θ inconnu et soit Θ_0 l'ensemble des paramètres correspondant à un écho suffisamment "grand". Le problème est alors de décider à partir de l'échantillon si oui ou non $\theta \in \Theta_0$, i.e. si oui ou non un missile est présent.

2.1.2 Définitions et Concepts.

1. Le problème d'un test statistique

Soit $\Theta_0 \subset \Theta$ et $\Theta_1 \subset \Theta$ tels que Θ_0 et Θ_1 forment une partition de Θ. Le problème d'un test statistique se résume de la façon suivante :

$$H_0 : \theta \in \Theta_0 \quad \textbf{\textit{versus}} \quad H_1 : \theta \in \Theta_1$$

où H_0 ***est l'hypothèse nulle*** et H_1 ***l'hypothèse alternative***. Chacune des hypothèse peut être de deux natures :

Pour *i=0* ou *i=1* si H_i correspond à un ensemble Θ_i réduit à un singleton $\{\theta_i\}$, H_i est dit simple. Dans le contraire, H_i est composite.

Etant donné l'hypothèse nulle $H_0 : \theta \in \Theta_0$, construire une procédure de test revient à construire à partir de l'échantillon ($X_1, X_2, ..., X_n$) une règle de décision φ_n qui indique si oui ou non H_0 est vérifiée. Formellement, on a la définition :

Définition. Un test ***simple*** est une fonction mesurable $\varphi_n : IR^n \to \{0;1\}$ qui ne dépend pas de θ. On accepte l'hypothèse nulle H_0 lorsque $\varphi_n = 0$ et on la rejette lorsque $\varphi_n = 1$, i.e. on accepte l'hypothèse alternative H_1.

Un test ***randomisé*** est une fonction mesurable : $\varphi_n : IR^n \to [0;1]$ qui ne dépend pas de θ. Lorsque $\varphi_n \in \{0;1\}$, les règles de décision sont les mêmes que pour les tests simples (si $\varphi_n = 0$ on accepte l'hypothèse nulle, si $\varphi_n = 1$ on la rejette). Lorsque $\varphi_n \in]0;1[$, on rejette l'hypothèse nulle avec la probabilité φ_n et on l'accepte donc avec probabilité $1 - \varphi_n$.

Un test simple φ_n est une ***variable aléatoire*** ne prenant que deux valeurs, 0 ou 1, c'est donc une variable de Bernoulli. On appelle zone de rejet ou zone critique ou région critique du test l'ensemble $R = \{\varphi_n((X_1, ..., X_n)) = 1\}$, i.e. la zone des observations qui conduisent à rejeter l'hypothèse nulle. Evidemment, construire un test simple est équivalent à donner une zone de rejet R alors le test s'écrit de manière unique $\varphi_n((X_1, ..., X_n)) = \mathbf{1}_R(X_1, ..., X_n)$.

Remarque. Par définition d'une statistique exhaustive $T(X_1, X_2, ..., X_n)$, elle contient toute l'information de l'échantillon permettant d'inférer sur θ. On recherche donc une zone R de rejet de H_0 sous la forme $R = \{T(X_1, ..., X_n) \in C\}$ pour un ensemble C à déterminer.

2. Risque des tests

Ayant construit un test, on prend la décision d'accepter ou non H_0 à partir de l'échantillon observé ($X_1, X_2, ..., X_n$). Il y a quatre possibilités :

– On accepte à raison H_0, i.e. $\varphi_n((X_1, ..., X_n)) = 0$ et $\theta \in \Theta_0$,
– On rejette à raison H_0, i.e. $\varphi_n((X_1, ..., X_n)) = 1$ et $\theta \in \Theta_1$,
– On rejette à tort H_0, i.e. $\varphi_n((X_1, ..., X_n)) = 1$ et $\theta \in \Theta_0$,
– On accepte à tort H_0, i.e. $\varphi_n((X_1, ..., X_n)) = 0$ et $\theta \in \Theta_1$.

On parlera dans les deux derniers cas ***d'erreurs de tests*** lié au fait qu'on prend une décision sur le paramètre θ inconnu à partir des observations ($X_1, X_2, ..., X_n$) uniquement. Rejeter à tort H_0 correspond à ***l'erreur de première espèce*** et accepter à tort H_0 ***l'erreur de seconde espèce***.

Le but du statisticien est de construire un test qui conduit à une erreur dans le moins de cas possibles.

On appelle erreur de première espèce ou erreur de type I, notée α, la probabilité de rejeter H_0 alors qu'elle est vraie. L'erreur α est aussi appelé niveau ou seuil de signification. Ainsi, $\alpha = P(\textit{accepter}\, H_1 \textit{ sachant que } H_0 \textit{ vraie})$ soit $\alpha = P\left(R/H_0\right)$.

On appelle erreur de deuxième espèce ou erreur de type II, notée β, la probabilité d'accepter H_0 alors qu'elle est fausse. Ainsi, $\beta = P(\textit{accepter}\, H_0 \textit{ sachant que } H_1 \textit{ vraie})$ soit $\beta = P\left(R^c/H_1\right)$, R^c désignant le complémentaire de *R*.

On appelle puissance du test $(1-\beta)$ pour H_1 la probabilité de retenir H_1 alors qu'elle est vraie soit $1-\beta = P\left(R/H_1\right)$.

Dans le cas où l'hypothèse alternative est composite ($\theta \in \Theta_1$), la puissance du test $1-\beta$ est fonction de θ et est appelée la **fonction puissance du test**.

Une fois que l'on a fixé raisonnablement α, il faut choisir une **variable de décision**, qui doit apporter le maximum d'informations sur le problème posé, et dont la loi sera différente selon que H_0 ou H_1 est vraie. La loi sous H_0 doit être connue. On définit alors la **région critique** *R* qui est l'ensemble des valeurs de la variable de décision qui conduisent à rejeter H_0 au profit de H_1. Sa forme est déterminée par la nature de H_1, et sa détermination exacte est donnée par $\alpha = P\left(R/H_0\right)$. La **région d'acceptation de** H_0 est son complémentaire R^c. Les points de jonction entre les deux régions sont ***les points critiques***.

Le choix d'un test sera donc le résultat d'un compromis entre risque de premier espèce et puissance du test. Le but du statisticien est donc de construire un test dont les risques de première et seconde espèce sont les plus faibles possibles (ou de manière équivalente un test dont le risque de première espèce est faible et la puissance est forte).

En résumé :

Vérité / *Décision*	H_0	H_1
H_0	$1-\alpha = P\left(R^c/H_0\right)$ *Niveau de confiance du test* *(Bonne décision)*	$\beta = P\left(R^c/H_1\right)$ *Erreur de deuxième espèce* *(Erreur de type II)* *(Mauvaise décision)*
H_1	$\alpha = P\left(R/H_0\right)$ *Erreur de première espèce* *(Erreur de Type I)* *(Mauvaise décision)*	$1-\beta = P\left(R/H_1\right)$ *Puissance du test* *(Bonne décision)*

2.2 L'approche de Neyman-Pearson et recherche de test uniformément plus puissant

Le principe de Neyman est de se fixer un seuil de tolérance sur le risque de première espèce appelé niveau.

Définition 1. On dit qu'un test est de niveau $\alpha \in [0;1]$ si son risque de première espèce est inférieur ou égal à α. Parmi les tests d'un niveau α fixé, il faut ensuite choisir celui qui a la plus grande puissance β, i.e. le plus petit risque de second espèce.

Définition 2. Soit R un sous-ensemble de l'espace échantillonnal. Nous qualifions R de meilleure région critique de taille α pour tester l'hypothèse simple $H_0: \theta = \theta_0$ contre l'hypothèse alternative simple $H_1: \theta = \theta_1$ si

(a) $P_{\theta_0}(X \in R) = \alpha$ et pour chaque sous-ensemble A de l'espace échantillonnal (b) $P_{\theta_0}(X \in A) = \alpha \implies P_{\theta_1}(X \in R) \geq P_{\theta_1}(X \in A)$

Définition 3. Soit $\alpha \in [0;1]$ et soit un test φ_n de niveau α et de région critique R. Le test φ_n est dit sans biais si $P_\theta(X \in R) \geq \alpha$ pour tout $\theta \in \Theta_1$.

Soit R de meilleure région critique de taille α pour tester l'hypothèse simple $H_0: \theta = \theta_0$ contre l'hypothèse alternative simple $H_1: \theta = \theta_1$. Notons $\gamma_R(\theta_1) = P_{\theta_1}(X \in R)$ la puissance du test c'est-à-dire la probabilité de rejeter H_0 alors qu'elle est fausse. Alors, $\gamma_R(\square_1) \geq \alpha$. Il est uniformément plus puissant (***UPP***) si pour tout test φ_n' de niveau α et de puissance $\gamma_R'(\theta)$, on a $\gamma_R(\theta) \geq \gamma_R'(\theta)$ pour tout $\theta \in \Theta_1$.

Le principe de Neyman consiste donc à trouver le test ***UPP*** pour un niveau α qui est fixé par le statisticien.

La fonction de vraisemblance est-elle differente selon l'hypothese en question?

Théorème de Neyman et Pearson (1933) : Hypothèses simples

On se propose de tester les hypothèses simples suivantes : $H_0: \theta = \theta_0$ ***versus*** $H_1: \theta = \theta_1$, $\theta_0 < \theta_1$. La fonction de densité et la vraisemblance pour θ en H_0 et H_1 sont respectivement données par $f_X(x, \theta_0)$ et $f_X(x, \theta_1)$, $L(x, \theta_0)$ et $L(x, \theta_1)$ en utilisant un test ayant pour région critique ***R*** qui satisfait pour $\forall k \geq 0$

$$\frac{L(x,\theta_0)}{L(x,\theta_1)} \leq k, \forall x \in R, \quad \frac{L(x,\theta_0)}{L(x,\theta_1)} \geq k\ , \forall x \in R^c \ (1) \ et \ \alpha = P(X \in R/H_0) \ (2)$$

Alors tout test satisfaisant (1) et (2) est un test ***UPP*** de niveau α. La région critique R est la meilleure région critique (optimale) définie par les points $x = (x_1, x_2, ..., x_n)$ vérifiant $R = \left\{x : \frac{L(x,\theta_0)}{L(x,\theta_1)} > k_\alpha\right\}$. La constante k_α dépend de α et est telle que $\alpha = P(X \in R/H_0)$

Démonstration.

Dans l'optique de Newman-Pearson, on considère que l'erreur la plus grave consiste à rejeter à tort l'hypothèse nulle. On fixe alors un seuil maximum α_0 au risque de première espèce et on cherche un test qui minimise le risque de seconde espèce. La région critique R, selon cette démarche, est déterminée sous la condition que $1 - \gamma_R(\theta)$ soit maximale sous la contrainte $\alpha \leq \alpha_0$.

Remarque. Il n'existe pas de test *UPP* pour $H_0: \theta = \theta_0$ ***versus*** $H_1: \theta \neq \theta_0$.

Exemple

Soit ($X_1, X_2, ..., X_n$) un échantillon issu d'une loi normale de variance σ^2 connue et d'espérance $\boldsymbol{\mu}$ inconnue. On désire tester: $H_0 : \mu = \mu_0$ ***versus*** $H_1 : \mu = \mu_1$ en supposant $\mu_0 < \mu_1$. La vraisemblance de l'échantillon est donnée par : $L(x,\mu) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$ et le rapport de vraisemblance par : $\frac{L(x,\mu_1)}{L(x,\mu_0)} = \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(\mu_1-\mu_0)x_i - \frac{n}{2\sigma^2}(\mu_1^2-\mu_0^2)\right)$. Ainsi, $\frac{L(x,\mu_1)}{L(x,\mu_0)} > c_\alpha$ est équivalent à $\bar{x}_n > \frac{\sigma^2 \ell n(c_\alpha)}{n(\mu_1-\mu_0)} + \frac{\mu_1+\mu_0}{2} = C$ où C est la constante telle que $P\left(x \in R \middle/ H_0\right) = P\left(\bar{x}_n > C \middle/ H_0\right) = \alpha$. En effet, comme $X_i \sim iidN(\mu,\sigma^2) \Rightarrow \bar{X}_n \sim iidN(\mu,\sigma^2/n)$. Ainsi, $P(\bar{x}_n > C) = P(Z > z_{1-\alpha})$ avec $z_{1-\alpha} = \frac{C-\mu_0}{\sigma/\sqrt{n}}$ et donc $C = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$. Par conséquent, la région critique optimale est alors donnée par : $R = \left\{x : \bar{x}_n > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}\right\}$. La variable $\bar{x}_n$ porte le nom de variable de décision. De manière générale, la variable de décision est une variable Z_n telle que $P\left(x \in R \middle/ H_0\right) = P\left(Z_n > C \middle/ H_0\right) = \alpha$

On peut de même déterminer $R = \left\{x : \frac{L(x,\theta_1)}{L(x,\theta_0)} < c_\alpha\right\}$ sous $H_0 : \mu = \mu_0$ ***versus*** $H_1 : \mu = \mu_1$ en supposant $\mu_0 > \mu_1$.

Test UPP : Hypothèses composites

2.3 Tests du Ratio de vraisemblance

On se propose de tester les hypothèses simples suivantes : $H_0: \theta = \theta_0$ ***versus*** $H_1: \theta \neq \theta_0$. La fonction de densité et la vraisemblance pour θ sont respectivement données par $f_X(x,\theta)$ et $L(x,\theta)$. On note le logarithme de la vraisemblance par $\mathcal{L}(x,\theta)$.

Définition. On définit le ratio (rapport) de vraisemblance en $H_0: \theta = \theta_0$ par, $\hat{\theta}_n^{MV}$ étant l'estimateur du maximum de vraisemblance :

$$\Lambda = \frac{L(x,\theta_0)}{L\left(x,\hat{\theta}_n^{MV}\right)}, \ \Lambda \leq 1.$$

On peut voir que $\Lambda \longrightarrow 1$ si H_0 est vraie alors que si c'est H_1 qui est vraie, $\Lambda \longrightarrow 0$. Dans ce cas, une règle de décision intuitive pour un risque α donné est celle-ci :

$$Rejeter\ H_0\ \ en\ faveur\ de\ H_1\ si\ \ \Lambda \leq c\ \ o\grave{u}\ c\ est\ tel\ que\ \alpha = P_{\theta_0}(\Lambda \leq c):$$

C'est le test du Ratio de Vraisemblance (Likelihood Ratio Test).

Exemple. Soit $X \sim N(\theta, 1)$. On se propose de tester $H_0: \theta = 0$ ***versus*** $H_1: \theta = 1$. Déterminer la région critique R et la constante c telle que $\Lambda \leq c$ $o\grave{u}\ c\ est\ tel\ que\ \alpha = P_{\theta_0}(\Lambda \leq c)$.

$$X \sim N(\theta, 1) \Longrightarrow f_X(x, \theta) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{(x-\theta)^2}{2}\right\}$$

Ainsi,

$$\Lambda = \frac{L(x, \theta_0 = 0)}{L(x, \theta_1 = 1)} = \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left\{-\sum_{i=1}^n \left(\frac{x_i^2}{2}\right)\right\}}{\left(\frac{1}{\sqrt{2\pi}}\right)^n exp\left\{-\sum_{i=1}^n \left(\frac{(x_i-1)^2}{2}\right)\right\}} \Longrightarrow$$

$$\Lambda = \frac{L(x, \theta_0 = 0)}{L(x, \theta_1 = 1)} = exp\left\{-\sum_{i=1}^n x_i + \frac{n}{2}\right\}.$$

Si $k>0$, l'ensemble de tous les points $(x_1, x_2, \dots, x_n)$ tels que

$$exp\left\{-\sum_{i=1}^n x_i + \frac{n}{2}\right\} > k \Leftrightarrow \sum_{i=1}^n x_i \geq \frac{n - 2logk}{2} \equiv c$$

Par suite,

$$R = \left\{(x_1, x_2, \dots, x_n) : \sum_{i=1}^n x_i \geq c\right\}$$

Ou de manière équivalente

$$R = \left\{(x_1, x_2, \dots, x_n) : \bar{x}_n \geq \frac{c}{n} \equiv c_1\right\}$$

Sous $H_0: \theta = 0$, $\bar{X}_n \sim N(0, 1/n)$. Le risque de première espèce $\alpha = P_{H_0}(\Lambda \leq c) = P_{H_0}(\bar{X}_n \geq c_1)$. La puissance du test $\gamma_R(\theta = 1) = P_{H_1}(\bar{X}_n \geq c_1)$ est telle que

$$\gamma_R(\theta = 1) = P_{H_1}(\bar{X}_n \geq c_1) = \int_{c_1}^{+\infty} \frac{1}{\sqrt{1/n}\sqrt{2\pi}} exp\left\{-\frac{(\bar{x}_n - 1)^2}{2(1/n)}\right\} d\bar{x}_n.$$

Par exemple, pour n=25, $\alpha = 0.05 \implies c_1 = 0.329$, $\gamma_R(\theta = 1) = 0.9996$.

Théorème. En supposant que les hypothèses suivantes sont vérifiées :

H0 : $\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$.
H1 : $\Theta \in IR^d$.
H2 : Le support $\{x : f(x, \theta) > 0\}$ ne dépend pas de θ.
H3 : Pour tout x la fonction $\theta \rightarrow f(x, \theta)$ est au moins deux fois continûment dérivable sur Θ
H4 : Pour tout $A \in \mathcal{A}$, l'intégrale $\int_A f(x, \theta)dx$ est deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation.

H5. Il existe une constante *c* et une fonction $M(x)$ telle que $\left|\frac{\partial^3 logf(x,\theta)}{\partial\theta^3}\right| \leq M(x)$ où $E[M(x)] < +\infty$ et $\theta_0 - c < \theta < \theta_0 + c$

Sous H_0: $\theta = \theta_0, \chi^2_{\mathrm{L}} = -2log(\Lambda) \xrightarrow{\mathcal{L}} \chi^2_1$

Démonstration

DEMO SA DWE NAN KRAZE AK TOUT DETAY

On note cette statistique par $\chi^2_{\mathrm{W}} = \left[\sqrt{n}\left(\hat{\theta}_n^{MV} - \theta_0\right)\sqrt{I_n(\theta_0)}\right]^2 \xrightarrow{\mathcal{L}} \chi^2_1$. Cette statistique déduite du ratio de vraisemblance porte le nom de statistique de Wald. Une règle de décision pour le test du ratio de vraisemblance basée sur la statistique de Wald (test de Wald) est celle-ci : $Rejeter\ H_0\ \ en\ faveur\ de\ H_1\ si\ \ \chi^2_{\mathrm{W}} \geq \chi^2_1$.

On peut aussi utiliser le test du score (test du score de Rao). En effet, $\frac{1}{\sqrt{n}}\mathcal{L}'(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\partial logf(X_i,\theta_0)}{\partial\theta}\right) \xrightarrow{\mathcal{L}} N\left(0, I_n(\theta)\right)$ où $S(\theta_0) = \left(\frac{\partial logf(X_1,\theta_0)}{\partial\theta}, \frac{\partial logf(X_2,\theta_0)}{\partial\theta}, \ldots, \frac{\partial logf(X_n,\theta_0)}{\partial\theta}\right)'$ (le vecteur score). Par suite,

$$\chi^2_{\mathrm{R}} = \left(\frac{\mathcal{L}'(\theta_0)}{\sqrt{I_n(\theta_0)}}\right)^2 \approx \chi^2_{\mathrm{W}} \xrightarrow{\mathcal{L}} \chi^2_1$$

• Le concept de p-valeur

En pratique, plutôt que de calculer la région critique en fonction de α, on préfère donner un seuil critique α^*, appelée ***p-valeur***, qui est la plus grande valeur de α conduisant à ne pas rejeter H_0. Cette information permet au lecteur de conclure à l'acceptation de H_0 pour tout risque de première espèce $\alpha \leq \alpha^*$, et à son rejet pour tout $\alpha > \alpha^*$.

2.3 Dualité entre intervalle de confiance et test statistique

Il existe une correspondance étroite entre un intervalle de confiance et un test d'hypothèses statistiques. Le théorème suivant établit cette dualité.

Théorème. Pour tout $\theta_0 \in \Theta$, on définit $R(\theta_0)$ la région d'acceptation d'un test de niveau α sous $H_0 : \theta = \theta_0$. Pour chaque $x \in D_X$, on définit dans l'espace des paramètres, un ensemble $C(x) = \{\theta_0 : x \in R(\theta_0)\}$. Alors l'ensemble aléatoire *C(X)* est un intervalle de confiance. Inversement, soit *C(X)* un intervalle de confiance de niveau $1-\alpha$. Si pour tout $\theta_0 \in \Theta$, on définit $R(\theta_0) = \{x : \theta_0 \in C(x)\}$, alors $R(\theta_0)$ est la région d'acceptation d'un test de niveau α sous $H_0 : \theta = \theta_0$.

Démonstration

a) On suppose que $R(\theta_0)$ est la région d'acceptation d'un test de niveau α sous $H_0 : \theta = \theta_0$. Par conséquent, $P_{\theta_0}(X \notin R(\theta_0)) \leq \alpha$ et donc $P_{\theta_0}(X \in R(\theta_0)) \geq 1-\alpha$. Comme θ est arbitraire, on peut remplacer θ_0 par θ et on peut prendre $P_\theta(X \in R(\theta_0)) = P_\theta(\theta \in C(X)) \geq 1-\alpha$. Ce qui permet de conclure que *C(X)* est un intervalle de confiance de niveau $1-\alpha$.

b) On suppose que *C(X)* est un intervalle de confiance de niveau $1-\alpha$ sous $H_0 : \theta = \theta_0$. Par conséquent, $P_{\theta_0}(\theta_0 \in C(X)) \leq \alpha$. On peut alors définir une région $R(\theta_0)$ de non acceptation de H_0 alors qu'elle est vraie telle que $P_{\theta_0}(X \notin R(\theta_0)) = P_{\theta_0}(\theta \notin C(X)) \leq \alpha$. Ce qui démontre qu'il s'agit d'un test de niveau α.

Résumé de la démarche générale du mécanisme des tests statistiques

La démarche de construction d'un test est la suivante :

– choix de H_0 et H_1,
– On choisit en général le risque de type I, α (souvent donne dans l'énoncé).
– détermination de la variable de décision,
– allure de la région critique en fonction de H_1,
– calcul de la région critique en fonction de α,
– calcul de la valeur expérimentale de la variable de décision,
– conclusion : rejet ou acceptation de H_0.

2.4 La pratique des tests statistiques paramétriques basiques

2.4.1 Test de conformité

Généralités

Soit X une variable aléatoire dont la loi dépend d'un paramètre inconnu θ et θ_0 une valeur numérique. On désire tester $H_0 : \theta = \theta_0$. L'hypothèse alternatives H_1 peut être de trois types :

- $H_1 : \theta \neq \theta_0$: il s'agit d'un test bilatéral (hypothèses simples)
- $H_1 : \theta > \theta_0$: il s'agit d'un test unilatéral à droite (hypothèses composites)
- $H_1 : \theta < \theta_0$: il s'agit d'un test unilatéral à gauche (hypothèses composites)

La région critique R (zone de rejet de H_0) est donnée par $\alpha = P(\textit{accepter}\, H_1 \ \textit{sachant que}\ H_0 \ \textit{vraie})$ soit $\alpha = P\left(c \in R \,/\, \theta = \theta_0\right)$, *c* est une constante à déterminer.

Dans le cas des tests bilatéraux, $R =]-\infty, c_1[\cup]c_2, +\infty[$ et donc $R^c = [c_1, c_2]$ de telle sorte que $1-\alpha = P\left(c_1 \leq X \leq c_2 \,/\, \theta = \theta_0\right)$, c_1 et c_2 sont deux constantes à déterminer.

Dans le cas des tests unilatéraux à droite, $R =]c, +\infty[$ et donc $R^c =]-\infty, c]$ de telle sorte que $1-\alpha = P\left(X \leq c \,/\, \theta = \theta_0\right)$, *c* est une constante à déterminer.

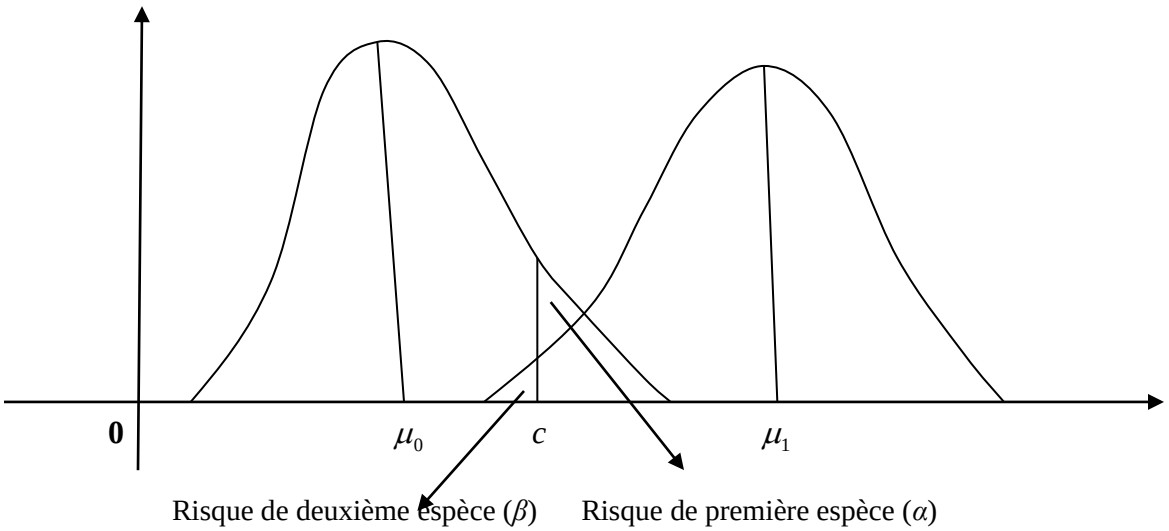
Dans le cas des tests unilatéraux à gauche, $R =]-\infty, c[$ et donc $R^c = [c, +\infty[$ de telle sorte que $1-\alpha = P\left(X \geq c \,/\, \theta = \theta_0\right)$, *c* est une constante à déterminer.

1. Test d'une moyenne

Hypothèse 1 : On suppose que l'échantillon ($X_1, X_2, ..., X_n$) provient d'une population normale c'est à dire $X_i \sim N(\mu, \sigma^2)$ et que σ est connu.

1.1 Test bilatéral. On se propose de tester : $H_0 : \mu = \mu_0$ ***versus*** $H_1 : \mu \neq \mu_0$. Or, $E(X) = \mu$ ce qui implique, on peut estimer μ par $\overline{X}_n$. Puisque $\mu \neq \mu_0$, la région de rejet de H_0 est donnée par $R =]-\infty, -c[\cup]c, +\infty[$ et donc la région de non rejet $R^c = [-c, c]$. On vérifie que $\overline{X}_n \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ sous H_0.

• **Hypothèses statistiques simples**. On choisit $\mu = \{\mu_0, \mu_1\}$ $\mu_0 < \mu_1$. On se propose de tester : $H_0 : \mu = \mu_0$ ***versus*** $H_1 : \mu = \mu_1$. Or, $E(X) = \mu$ ce qui implique, on peut μ par $\overline{X}_n$. Puisque $\mu_0 < \mu_1$, la région de rejet de H_0 est donnée par $R =]c, +\infty[$ et donc la région de non rejet $R^c =]-\infty, c]$. On vérifie que $\overline{X}_n \sim N\left(\mu_0, \frac{\sigma^2}{n}\right)$ sous H_0.



Ainsi, le risque de première espèce α est telle que $\alpha \geq P\left(\left|\overline{X}_n\right| \geq c \Big/ \mu = \mu_0\right)$ tout en cherchant à minimiser $\beta = P\left(\left|\overline{X}_n\right| < c \Big/ \mu = \mu_1\right)$. Or, $P\left(\left|\overline{X}_n\right| \geq c \Big/ \mu = \mu_0\right) = P\left(\left|\frac{\overline{X}_x - \mu}{\sigma/\sqrt{n}}\right| \geq \frac{c - \mu_0}{\sigma/\sqrt{n}} \Big/ \mu = \mu_0\right)$ et par conséquent,

$P\left(\left|\overline{X}_n\right| \geq c \Big/ \mu = \mu_0\right) \equiv \varphi_{n,0}(c) = 2\left(1 - F_{\overline{X}_n}\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right)\right) \leq \alpha/2$ et $P\left(\left|\overline{X}_n\right| < c \Big/ \mu = \mu_1\right) \equiv \varphi_{n,1}(c) = 2F_{\overline{X}_n}\left(\frac{c - \mu_1}{\sigma/\sqrt{n}}\right) - 1$

est minimum. Comme $\varphi_{n,1}(c)$ est croissante, on peut choisir c le plus petit que possible vérifiant l'inégalité et comme $\varphi_{n,0}(c)$ est décroissante, on peut alors choisir c tel que $1 - F_{\overline{X}_n}\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right) = \alpha/2$ soit $c = \mu_0 + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Ainsi,

$\beta \equiv P\left(\left|\overline{X}_n\right| < c \middle/ \mu = \mu_1\right) = 2F_{\overline{X}_n}\left(z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) - 1$. La puissance du test est d'autant plus grande que β est petit c'est-à-dire quand $\mu_1 - \mu_0$ est grand ou n est grand.

• **Hypothèses statistiques composites**. On choisit $\mu = [\mu_0, +\infty[$ On se propose de tester : $H_0 : \mu = \mu_0$ ***versus*** $H_1 : \mu > \mu_0$.

On rejette H_0 que si $\overline{X}_n \geq c$ et donc $c = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$, $\alpha \equiv P\left(\overline{X}_n \geq c \middle/ H_0\right) = 1 - F_{\overline{X}_n}\left(\frac{c - \mu_0}{\sigma/\sqrt{n}}\right)$ et $\beta \equiv P\left(\overline{X}_n < c \middle/ H_1\right) = F_{\overline{X}_n}\left(\frac{c - \mu_1}{\sigma/\sqrt{n}}\right) = F_{\overline{X}_n}\left(z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$. **La courbe d'efficacité** est le graphique de la fonction $\varepsilon(\mu) \equiv P\left(\overline{X}_n < c \middle/ H_1\right) = F_{\overline{X}_n}\left(\frac{c - \mu_1}{\sigma/\sqrt{n}}\right)$.

Hypothèse 2 : On suppose que l'échantillon ($X_1, X_2, ..., X_n$) provient d'une population normale c'est à dire $X_i \sim N(\mu, \sigma^2)$ et que σ est inconnu.

Si σ est inconnu, on l'estime par $s_n^2 = \frac{1}{n-1}\sum_{1=1}^{n}(X_i - \overline{X})^2$. Le rapport $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0;1)$ devient $\frac{\overline{X}_n - \mu}{s_n/\sqrt{n}} \sim t_{n-1}$. De ce fait, on décide de rejeter l'hypothèse nulle que si $\overline{X}_n \geq c = \mu_0 + t_{n-1,1-\alpha}\frac{s_n}{\sqrt{n}}$. Evidemment, pour n suffisamment élevé, on peut toujours accepter que $\frac{\overline{X}_n - \mu}{s_n/\sqrt{n}} \xrightarrow{ass} N(0;1)$ et donc le rejet de l'hypothèse nulle est telle que $\overline{X}_n \geq c \approx \mu_0 + z_{1-\alpha}\frac{s_n}{\sqrt{n}}$.

2. Test sur la variance d'une population

On suppose que l'échantillon ($X_1, X_2, ..., X_n$) provient d'une population normale c'est à dire $X_i \sim N(\mu, \sigma^2)$. On désire alors tester $H_0 : \sigma^2 = \sigma_0^2$ *versus*

a) $H_1 : \sigma^2 \neq \sigma_0^2$ ((hypothèses simples)

○ **On suppose que μ est connu.** Dans ce cas, la meilleure estimation de la variance est donnée par $S_n^2 = \frac{1}{n}\sum_{1=1}^{n}(X_i - \mu)^2$. Dans ce cas, sous l'hypothèse H_0, le rapport $\frac{nS_n^2}{\sigma_0^2} \sim \chi_n^2$ (statistique du test). On rejette alors H_0 si $S_n^2 < \frac{\sigma_0^2}{n}\chi_{n,\alpha/2}^2$ ou $S_n^2 > \frac{\sigma_0^2}{n}\chi_{n,1-\alpha/2}^2$ ou de manière équivalente si $\frac{nS_n^2}{\sigma_0^2} < \chi_{n,\alpha/2}^2$ ou $\frac{nS_n^2}{\sigma_0^2} > \chi_{n,1-\alpha/2}^2$

○ **On suppose que μ est inconnu.** Dans ce cas, on estime μ par $\overline{X}_n$ et la meilleure estimation de la variance est donnée par $s_n^2 = \frac{1}{n-1}\sum_{1=1}^{n}(X_i - \overline{X}_n)^2$. Dans ce cas, sous l'hypothèse H_0, le rapport $\frac{(n-1)s_n^2}{\sigma_0^2} \sim \chi^2_{n-1}$ (statistique du test). On rejette alors H_0 si $s_n^2 < \frac{\sigma_0^2}{n-1}\chi^2_{n-1,\alpha/2}$ ou $s_n^2 > \frac{\sigma_0^2}{n-1}\chi^2_{n-1,1-\alpha/2}$ ou de manière équivalente si $\frac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{n-1,\alpha/2}$ ou $\frac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha/2}$

b) $H_1 : \sigma^2 > \sigma_0^2$ (hypothèses composites, test unilatéral à droite).

Si μ est connu, on rejette H_0 que si $\frac{nS_n^2}{\sigma_0^2} > \chi^2_{n,1-\alpha}$. Par contre, si μ est inconnu, on rejette H_0 que si $\frac{(n-1)s_n^2}{\sigma_0^2} > \chi^2_{n-1,1-\alpha}$.

c) $H_1 : \sigma^2 < \sigma_0^2$ ((hypothèses composites, test unilatéral à gauche).

Si μ est connu, on rejette H_0 que si $\frac{nS_n^2}{\sigma_0^2} < \chi^2_{n,1-\alpha}$. Par contre, si μ est inconnu, on rejette H_0 que si $\frac{(n-1)s_n^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha}$.

3. Test sur une proportion

Dans la population étudiée, une proportion ***p*** des individus possèdent un certain caractère *C*. On se propose de comparer cette proportion ***p*** à une valeur de référence p_0 ($H_0 : p = p_0$). On considère un échantillon d'individus de taille n de cette population. La variable aléatoire X_i, égale à 1 si l'individu i possède le caractère *C*, suit une loi de Bernoulli $\mathcal{B}(p)$, et le nombre d'individus $\sum_{i=1}^{n} X_i$ possédant ce caractère suit une loi binomiale $\mathcal{B}(n, p)$.

Si n est suffisamment grand, de sorte que $np > 5$ et $n(1-p) > 5$, on peut considérer (loi des grands nombres) que $\sum_{i=1}^{n} X_i \xrightarrow{n\to\infty} N(np; np(1-p))$ d'où la fréquence empirique $\left(\wp_n = (1/n)\sum_{i=1}^{n} X_i\right) \xrightarrow{n\to\infty} N\left(p, \frac{p(1-p)}{n}\right)$.

Si n est trop petit, le test est construit sur la loi binomiale, et on peut utiliser les *abaques*.

a) $H_1 : p \neq p_0$ (hypothèses simples). Sous H_0, $\wp_n \xrightarrow{n\to\infty} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$. On rejette H_0 que si $\left|\hat{\wp}_n - p_0\right| > z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$ ou de manière équivalente si $\frac{\left|\hat{\wp}_n - p_0\right|}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/2}$

b) $H_1 : p > p_0$ (hypothèses composites, test unilatéral à droite). On rejette H_0 que si $\hat{\wp}_n > -z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} + p_0$ ou de manière équivalente si $\frac{\hat{\wp}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > -z_\alpha$

c) $H_1 : p > p_0$ (hypothèses composites, test unilatéral à gauche). On rejette H_0 que si $\hat{\wp}_n < z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} + p_0$ ou de manière équivalente si $\frac{\hat{\wp}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_\alpha$

Exemples

1. Un vétérinaire du quartier affirme que la durée de vie des chats siamois du quartier est de 11.5 ans avec un écart type de 2.4 ans. Pour vérifier cette affirmation, quelqu'un prélève un échantillon indépendant de 60 chats siamois et note leur durée de vie. On a trouve une durée de vie moyenne de 12 ans.

1.1 En suivant les différentes étapes d'un test d'hypothèses, cette affirmation est-elle fondée ? On donne $\alpha = 0.05$.

1.2 Calculer la probabilité de commettre une erreur de deuxième espèce pour une contre hypothèse $\mu_1 = 12.5$

1.1 Construire un tableau donnant la puissance du test en fonction de μ_1 pour des valeurs allant de μ_1=11.0 à μ_1=12.0 par des multiples de 0.1. Tracer la courbe de puissance du test.

Réponses

1.1 a) $H_0 : \mu = 11.5$ *versus* $H_1 : \mu \neq 11.5$

b) On a $\sigma^2 = (2.4)^2$ et n=60. On s'intéresse à l'estimateur $\overline{X}_n$. Puisque σ^2 est connue et que n>30, on peut admettre que $Z_n = \frac{\overline{X}_n - 11.5}{\sqrt{\frac{(2.4)^2}{60}}} \xrightarrow{ass} N(0;1)$. Le test est bilatéral avec une probabilité de commettre une erreur de première espèce $\alpha = 0.05$. La région critique R est telle que $P(|Z_n| > z_{0.05/2}) = 0.05/2 = 0.025$ $\Leftrightarrow Z_n < -1.96$ ou $Z_n > 1.96$. La région d'acceptation est donc $-1.96 \le Z_n \le 1.96$ ou de manière équivalente $10.89 \le \overline{X}_n \le 12.11$. On a : $Z_n = 1.6139 \Rightarrow -1.96 \le Z_n \le 1.96$ et donc on ne peut rejeter $H_0 : \mu = 11.5$.

1.1 On a $\beta = P\left(R^c / H_1\right) =$ $P(\textit{accepter}\, H_0 \,\textit{sachant qu'elle est fausse})$. Il revient alors de calculer

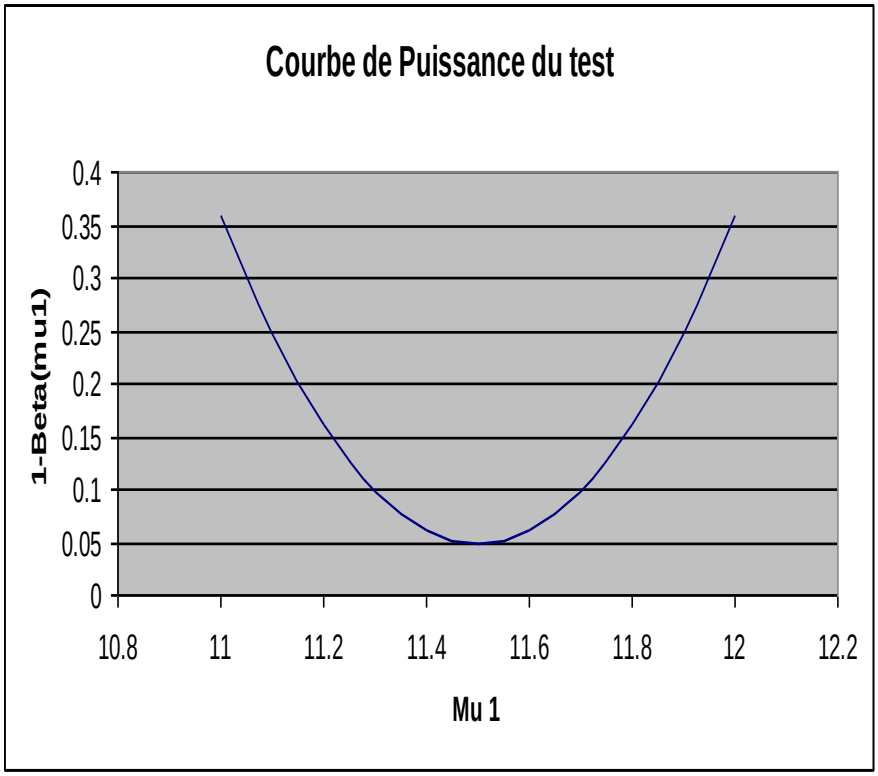
$$\beta(12.5) = P\left(10.89 \le \overline{X}_n \le 12.11 \,/\, \mu_1 = 12.5\right) \Rightarrow \qquad \beta(12.5) = P\left(\frac{10.89 - 12.5}{0.3098} \le Z_n \le \frac{12.11 - 12.5}{0.3098}\right) = 0.1038.$$

C'est la probabilité d'accepter que $\mu = 11.5$ alors qu'elle est en réalité 12.5. Ainsi, $1 - \beta(12.5) = 0.8962$ qui est la puissance du test pour une contre hypothèse $\mu_1 = 12.5$ c'est-à-dire c'est la probabilité de détecter que la moyenne n'est pas 11.5 quand elle est en réalité 12.5.

1.2 En procédant comme en 1.2, on a les résultats suivants :

Contre Hypothèse : μ_1	Erreur de deuxième espèce : $\beta(\mu_1)$	Puissance du test : $1-\beta(\mu_1)$
11.0	0.6404	0.3596
11.1	0.7512	0.2488
11.2	0.8397	0.1603
11.3	0.9021	0.0979
11.4	0.9395	0.0605
11.5	0.9512	0.0488
11.6	0.9395	0.0605
11.7	0.9021	0.0979
11.8	0.8597	0.1603
11.9	0.7512	0.2488
12.0	0.6404	0.3596

Courbe de puissance du test :



2. Soit $X \sim N(\mu, \sigma^2 = 16)$. Disposant un échantillon indépendant issu de cette population normale, on veut choir entre les deux hypothèses suivantes : $H_0 : \mu = 2$ *versus* $H_1 : \mu = 3$.

a) Résoudre ce problème par la méthode de Newman et Pearson
b) Dans le cas où n=100 et $\alpha = 0.05$, calculer la puissance de ce test. Que conclure ?
c) Quelle doit être la taille de l'échantillon minimum n_0 pour que la puissance soit supérieure à 0.95 ? 0.99 ?

Réponses

a) La vraisemblance s'écrit : $L(x_1,..x_n;\mu/\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$. Du théorème de NP, on déduit la forme de la région critique $\frac{L_0}{L_1}\le k$: $-\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i-2)^2-(x_i-3)^2]\le \ell nk \Leftrightarrow \sum_{i=1}^{n}x_i\ge \sigma^2\ell nk+5/2\Leftrightarrow \overline{X}_n\ge C$, $C=\frac{\sigma^2\ell nk}{n}+\frac{5}{n}$. La région critique est donnée par $R=\{(x_1,...,x_n)/\overline{X}_n\ge C\}$ avec $\alpha=P(\overline{X}_n\ge C/\mu=2)$ de telle sorte que $C=2+z_\alpha\frac{2}{\sqrt{n}}$ ($\sigma=2$) et donc $R=\{(x_1,...,x_n)/\overline{X}_n\ge 2+z_\alpha\frac{2}{\sqrt{n}}\}$

b) Pour $\alpha=0.05$, $z_\alpha=1.6449$ et $C=2.33$ et donc $R=\{(x_1,...,x_{100})/\overline{X}_{100}\ge 2.33\}$. La puissance est $1-\beta(\mu=3)=P(\overline{X}_{100}\ge 2.33/\mu=3)=P\left(Z\ge\frac{2.33-3}{2/\sqrt{100}}\right)$ $\Rightarrow$ $1-\beta(\mu=3)=P(Z\ge -3.35)=0.996$ et donc $\beta(\mu=3)=0.004$. Dans ce cas $\beta(\mu=3)=0.004<\alpha=0.05$ ce qui est contraire à l'optique de NP qui a permis de construire ce test.

c) Pour que $1-\beta\equiv P(Z\ge c)\ge 0.95$, $c=\frac{C-3}{2/\sqrt{n}}=\frac{2+z_\alpha\frac{2}{\sqrt{n}}-3}{2/\sqrt{n}}=\frac{\frac{3.2898}{\sqrt{n}}-1}{2/\sqrt{n}}$ $\Rightarrow$ $\frac{\frac{3.2898}{\sqrt{n}}-1}{2/\sqrt{n}}\le -1.6449\Leftrightarrow$ $n_0=44$. Pour cette taille d'échantillon, on a $\beta=\alpha=0.05$ et C=2.5 à équidistance de $H_0:\mu=2$ et $H_1:\mu=3$. Dès que n sera inferieur à n_0, on aura $\beta<\alpha=0.05$ et donc il faudra diminuer le risque α pour rester cohérent avec l'optique de NP. La condition $1-\beta\equiv P(Z\ge c)\ge 0.99\Rightarrow$ $\frac{\frac{3.2898}{\sqrt{n}}-1}{2/\sqrt{n}}\le -2.3263$ et donc $n_0=64$. La taille ici étant n=100, il faut choisir $\alpha<0.01$ pour rester cohérent avec NP.

2.4.2. Test paramétrique de comparaison de paramètres issus de deux populations gaussiennes indépendantes

L'objectif de cette section est de dire si deux échantillons indépendants sont issus d'une même population ou non. Voici quelques exemples d'application :

– les rendements journaliers de deux usines d'un même groupe sont-ils semblables ?
– les ventes par semaine de deux actions sont-elles similaires ?

On formule le problème de la façon suivante : on observe deux échantillons $(X_{1,1},X_{1,2},.....,X_{1,n_1})$ et $(X_{2,1},X_{2,2},.....,X_{2,n_2})$ (observations appariées : deux observations sur une même variable), indépendants pour une même variable aléatoire *X* et de fonction de répartition respective $F_1(x)$ et $F_2(x)$, issus de deux populations indépendantes P_1 et P_2. Le test exact revient à tester l'égalité de ces fonctions de répartitions : $H_0:F_1(x)=F_2(x)$ contre $H_1:F_1(x)\ne F_2(x)$ mais en pratique, on se contente de tester l'égalité des (proportions) moyennes (μ_1,μ_2) et des variances (σ_1^2,σ_2^2) où μ_1 et μ_2 sont les moyennes, σ_1^2 et σ_2^2 les variances respectives de X_1 et X_2 au niveau des populations respectives P_1 et P_2. On suppose qu'au niveau de la population $X\sim N(\mu,\sigma^2)$.

1. Test de rapport de deux variances

On suppose que l'échantillon $(X_1, X_2, ..., X_n)$ provient d'une population normale c'est à dire $X_i \sim N(\mu, \sigma^2)$. On désire alors tester $H_0 : \sigma_1^2 = \sigma_2^2$ *versus*

a) $H_1 : \sigma_1^2 \neq \sigma_2^2$ (hypothèses simples, test bilatéral)

○ **On suppose que μ est connu.** Dans ce cas, la meilleure estimation de la variance est donnée respectivement par $S_{1n}^2 = \frac{1}{n_1}\sum_{1=1}^{n_1}(X_{1i} - \mu_1)^2$ pour la première population et $S_{2n}^2 = \frac{1}{n_2}\sum_{1=1}^{n_2}(X_{2i} - \mu_2)^2$ pour la deuxième population. Dans ce cas, sous l'hypothèse H_0, les rapports $\frac{n_1 S_{1n}^2}{\sigma_1^2} \sim \chi_{n_1}^2$ et $\frac{n_2 S_{2n}^2}{\sigma_2^2} \sim \chi_{n_2}^2$ de telle sorte que $F = \frac{n_1 S_{1n}^2 \big/ \sigma_1^2 n_1}{n_2 S_{2n}^2 \big/ \sigma_2^2 n_2} = \frac{S_{1n}^2}{S_{2n}^2} \sim F_{n_1, n_2}$ ($\sigma_1^2 = \sigma_2^2$) (statistique du test). On rejette alors H_0 si $F < F_{n_1,n_2}^{\alpha/2}$ ou $F > F_{n_1,n_2}^{1-\alpha/2}$.

○ **On suppose que μ est inconnu.** Dans ce cas, la meilleure estimation de la variance est donnée respectivement par $s_{1n}^2 = \frac{1}{n_1 - 1}\sum_{1=1}^{n_1}(X_{1i} - \overline{X}_{1n})^2$ pour la première population et $s_{2n}^2 = \frac{1}{n_2 - 1}\sum_{1=1}^{n_2}(X_{2i} - \overline{X}_{2n})^2$ pour la deuxième population. Dans ce cas, sous l'hypothèse H_0, les rapports $\frac{(n_1 - 1)s_{1n}^2}{\sigma_1^2} \sim \chi_{n_1 - 1}^2$ et $\frac{(n_2 - 1)s_{2n}^2}{\sigma_2^2} \sim \chi_{n_2 - 1}^2$ de telle sorte que $F = \frac{(n_1 - 1)s_{1n}^2 \big/ \sigma_1^2 (n_1 - 1)}{(n_2 - 1)s_{2n}^2 \big/ \sigma_2^2 (n_2 - 1)} = \frac{s_{1n}^2}{s_{2n}^2} \sim F_{n_1 - 1, n_2 - 1}$ ($\sigma_1^2 = \sigma_2^2$) (statistique du test). On rejette alors H_0 si $F < F_{n_1 - 1, n_2 - 1}^{\alpha/2}$ ou $F > F_{n_1 - 1, n_2 - 1}^{1-\alpha/2}$.

b) $H_1 : \sigma_1^2 > \sigma_2^2$ (hypothèses composites, test unilatéral à droite).

Si μ est connu, on rejette H_0 que si $F > F_{n_1,n_2}^{1-\alpha}$. Par contre, si μ est inconnu, on rejette H_0 que si $F > F_{n_1 - 1, n_2 - 1}^{1-\alpha}$.

c) $H_1 : \sigma_1^2 < \sigma_2^2$ (hypothèses composites, test unilatéral à gauche).

Si μ est connu, on rejette H_0 que si $F < F_{n_1,n_2}^{1-\alpha}$. Par contre, si μ est inconnu, on rejette H_0 que si $F < F_{n_1 - 1, n_2 - 1}^{1-\alpha}$.

2. Test de comparaison de moyennes de deux populations indépendantes

On considère deux variables aléatoires X_1 et X_2 indépendantes issues respectivement de deux populations P_1 et P_2 indépendantes. Soit μ_1 et μ_2 les moyennes, σ_1^2 et σ_2^2 les variances respectives de X_1 et X_2. On dispose d'un n_1-échantillon de X_1 qui donne une moyenne $\overline{X}_1$ et une variance s_1^2 et d'un n_2-échantillon de X_2 qui donne une moyenne $\overline{X}_2$ et une variance s_2^2. On désire tester $H_0 : \mu_1 = \mu_2$ *versus*

a) $H_1 : \mu_1 \neq \mu_2$ (hypothèses simples, test bilatéral)

° **Hypothèse 1** : On suppose que σ_1^2 et σ_2^2 sont connues. Sous H_0 , $Z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0;1)$. Au risque α , la zone de rejet (région critique) de H_0 est telle que $P(|Z| > z_{\alpha/2}) = \alpha$ et la région d'acceptation est donnée par $P(|Z| \leq z_{\alpha/2}) = 1-\alpha$.

° **Hypothèse 2** : On suppose que σ_1^2 et σ_2^2 sont inconnues.

i. Sous hypothèse 21 : on suppose que $\sigma_1^2 = \sigma_2^2$. Dans ce cas, la meilleure estimation de la variance est donnée respectivement par $s_{1n}^2 = \dfrac{1}{n_1 - 1}\sum_{1=1}^{n_1}(X_{1i} - \overline{X}_{1n})^2$ pour la première population et $s_{2n}^2 = \dfrac{1}{n_2 - 1}\sum_{1=1}^{n_2}(X_{2i} - \overline{X}_{2n})^2$ pour la deuxième population. On définit alors $s^2 = \dfrac{(n_1-1)s_{1n}^2 + (n_2-1)s_{2n}^2}{n_1 + n_2 - 2}$. Sous H_0, $t = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}} = \dfrac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim t_{n_1+n_2-2}$ car on avait supposé que $\sigma_1^2 = \sigma_2^2$. Au risque α , la zone de rejet (région critique) de H_0 est telle que $P(|t| > t_{n_1+n_2-2}^{\alpha/2}) = \alpha$ et la région d'acceptation est donnée par $P(|t| \leq t_{n_1+n_2-2}^{\alpha/2}) = 1-\alpha$. Evidemment pour des échantillons de grande taille, on peut toujours utiliser le fait que $t = \dfrac{\overline{X}_1 - \overline{X}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \xrightarrow{n_1, n_2 \to \infty} N(0;1)$.

ii. Sous hypothèse 22 : on suppose que $\sigma_1^2 = k\sigma_2^2$, $k > 1$. Sous H_0 , $t = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_{1n}^2}{n_1} + \dfrac{s_{2n}^2}{n_2}}} \sim t_\nu$ où ν est un entier naturel non nul tel que $\nu = \dfrac{\left(\dfrac{s_{1n}^2}{n_1} + \dfrac{s_{2n}^2}{n_2}\right)^2}{\dfrac{s_{1n}^4}{n_1^2(n_1-1)} + \dfrac{s_{2n}^4}{n_2^2(n_2-1)}}$ (Méthode de Welch). Au risque α , la zone de rejet (région critique) de H_0 est telle que $P(|t| > t_\nu^{\alpha/2}) = \alpha$ et la région d'acceptation est donnée par $P(|t| \leq t_\nu^{\alpha/2}) = 1-\alpha$. Dans le cas des échantillons de grande taille, on peut toujours utiliser le fait que $t = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{s_{1n}^2}{n_1} + \dfrac{s_{2n}^2}{n_2}}} \xrightarrow{n_1, n_2 \to \infty} N(0;1)$

b) $H_1 : \mu_1 > \mu_2$ (hypothèses composites, test unilatéral à droite). Au risque α , la zone de rejet (région critique) de H_0 est telle que $P(Z > z_\alpha) = \alpha$ et la région d'acceptation est donnée par $P(Z \leq z_\alpha) = 1-\alpha$.

c) $H_1 : \mu_1 < \mu_2$ (hypothèses composites, test unilatéral à gauche). Au risque α, la zone de rejet (région critique) de H_0 est telle que $P(Z < -z_\alpha) = \alpha$ et la région d'acceptation est donnée par $P(Z \geq -z_\alpha) = 1 - \alpha$.

3. Test de comparaison de proportions de deux populations indépendantes

Deux populations possèdent des individus ayant un certain caractère, en proportion p_1 et p_2. L'objet du présent test est de tester : $H_0 : p_1 = p_2$. On prélève deux échantillons *iid* $(X_{11}, X_{12}, ..., X_{1n_1})$ et $(X_{21}, X_{22}, ..., X_{2n_3})$ de tailles respectives n_1 et n_2 et soit les proportions échantillonnales respectives $\hat{p}_1$ et $\hat{p}_2$ d'individus ayant ce caractère. Les tailles sont supposées suffisamment grandes. Ainsi les lois des fréquences empirique $\hat{p}_1$ et $\hat{p}_2$ peuvent être approximées par des lois normales, d'où $\left(\hat{p}_{1n} = (1/n_1)\sum_{i=1}^{n_1} X_{i1} \right) \xrightarrow{n_1 \to \infty} N\left(p_1, \frac{p_1(1-p_1)}{n_1} \right)$ et $\left(\hat{p}_{2n} = (1/n_2)\sum_{i=1}^{n_2} X_{2i} \right) \xrightarrow{n_2 \to \infty} N\left(p_2, \frac{p_2(1-p_2)}{n_2} \right)$ pour n_1 et n_2 suffisamment grands Dans la pratique, on peut commencer à utiliser l'approximation à condition que $n_i p_i > 5$ et $n_i p_i (1-p_i) > 5$, *i=1,2*.

a) $H_1 : p_1 \neq p_2$ (hypothèses simples). Sous H_0, soit $p_1 = p_2 = p$, $\hat{p}_{1n} - \hat{p}_{2n} \xrightarrow{n_1, n_2 \to \infty} N\left(0, \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}\right)$. On estime p par $\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ et la statistique $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \xrightarrow{ass} N(0;1)$. On rejette H_0 que si $|\hat{p}_{1n} - \hat{p}_{2n}| > z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ ou de manière équivalente si $\frac{|\hat{p}_{1n} - \hat{p}_{2n}|}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > z_{\alpha/2}$.

Si n_1 et n_2 sont trop petits, le test est construit sur la loi binomiale, et on peut utiliser les *abaques*.

b) $H_1 : p_1 > p_2$ (hypothèses composites, test unilatéral à droite). On rejette H_0 que si $\hat{p}_{1n} - \hat{p}_{2n} > z_\alpha \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ ou de manière équivalente si $\frac{\hat{p}_{1n} - \hat{p}_{2n}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} > z_\alpha$.

c) $H_1 : p_1 < p_2$ (hypothèses composites, test unilatéral à gauche). On rejette H_0 que si $\hat{p}_{1n} - \hat{p}_{2n} < -z_\alpha \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ ou de manière équivalente si $\frac{\hat{p}_{1n} - \hat{p}_{2n}}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < -z_\alpha$.

2.4.3 Tests du chi-deux

2.4.3.1. Définition

Distribution multinomiale (rappel). On considère une expérience aléatoire générant un espace échantillon Ω ''***partitionné***'' en *r+1* sous-ensembles S_i (succès # *i, i=1,2,..,r*) et E (échec) de probabilité respective de réalisation p_i (i=1,2,..r) et $1-\sum_{i=1}^{r} p_i$ telles que $0 \le p_i \le 1$. ***On répète n fois de façon indépendante*** cette expérience aléatoire. On définit le *r-uplet* de variables aléatoires discrètes $(X_1, X_2, ..., X_r)$ ***''nombre de fois où l'on a simultanément*** S_1, S_2, …, S_r ''. La probabilité d'avoir $(X_1, X_2, ..., X_r)$ respectivement x_1 ***fois,*** x_2 fois …., x_r ***fois*** est donnée par :

$P\left(\bigcap_{i=1}^{r}(X_i = x_i)\right) \frac{n!}{\prod_{i=1}^{r} x_i!(n-\sum_{i=1}^{r} x_i)!} p_1^{x_1} p_2^{x_2} ....p^{x_r} \left(1-\sum_{i=1}^{r} p_i\right)^{n-\sum_{i=1}^{r} x_i}$ si $(x_1, x_2, ..., x_r) \in D_{X_1 ... X_r}$ et

$P\left(\bigcap_{i=1}^{r}(X_i = x_i)\right) = 0$ sinon où $D_{X_1 ... X_r} = \{(x_1, ..., x_r) \in S^r / \sum_{i=1}^{r} x_i \le n\}$, $S = \{0,1,2,....,n\}$

Le *r-uplet* de variables aléatoires discrètes $(X_1, X_2, ..., X_r)$ suit une loi multinomiale de paramètres *n*, $p_1, p_2, ..., p_r$. On note: $(X_1, X_2, ..., X_r) \sim \mathcal{M}.(n\,;\, p_1, p_2, ..., p_r)$. C'est donc une généralisation de la loi binomiale. On admet que $X_j \sim B(n, p_j)$ où $E(X_j) = np_j$ et $V(X_j) = np_j(1-p_j)$.

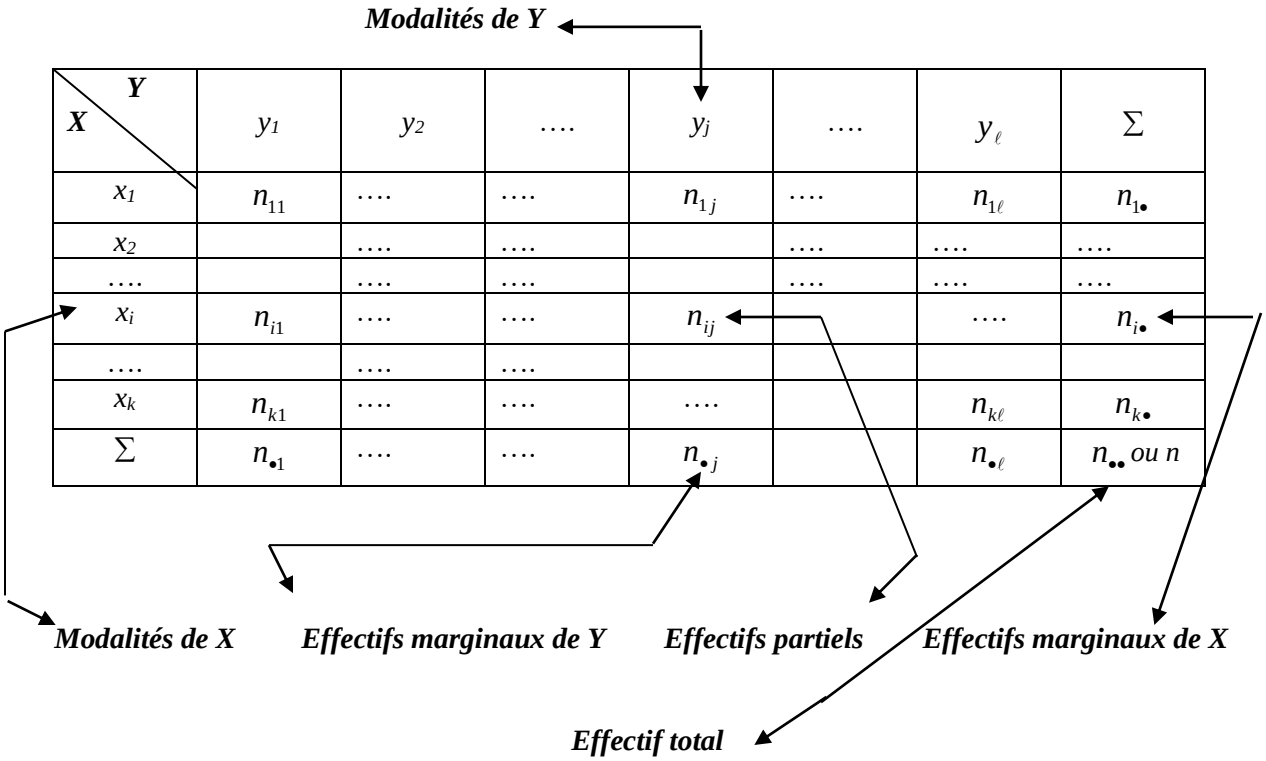
Théorème. La variable aléatoire $\sum_{j=1}^{r}\left(\frac{(X_j - np_j)^2}{np_j}\right) \xrightarrow{n\to+\infty} \chi^2_{r-1}$. Ce théorème se démontre de façon rigoureuse en utilisant le concept de fonction caractéristique d'une distribution d'un vecteur aléatoire. On peut tout de même remarquer que, en supposant np_j constant, $p_j \in]0\,;1[$ et $p_j \to 0$, $(X_j \sim B(n, p_j)) \xrightarrow{n\to\infty} P(\lambda = np_j)$ quand $n \to +\infty$ (théorème de Poisson) et par conséquent $\frac{X_j - np_j}{\sqrt{np_j}} \xrightarrow{n\to\infty} \mathcal{N}(0\,;1)$ (théorème limite centrale) en se rappelant que pour la loi de Poisson $\lambda = E(X_j) = V(X_j) = np_j$ et donc $\left(\frac{X_j - np_j}{\sqrt{np_j}}\right)^2 \xrightarrow{n\to\infty} \chi^2_1$. De ce fait, $\sum_{j=1}^{r}\left(\frac{(X_j - np_j)^2}{np_j}\right) \xrightarrow{n\to+\infty} \chi^2_{r-1}$ (*QED*).

2.4.3.2 Tests

1. Test d'indépendance

On rappelle que deux variables X et Y sont dites interdépendantes si elles varient conjointement, généralement influencées par un ou plusieurs facteurs externes quelconques. Si par contre, l'une ou l'autre de ces variables peut varier sans que l'autre n'en soit influencée, les variables X et Y sont dites alors indépendantes.

Supposons que X et Y soient qualitatives ou tout au moins traitées comme telles. On peut les représenter simultanément dans un tableau à double entrée appelée tableau de contingence (au moins une des variables est qualitative) .On considère que les deux variables statistiques *X* et *Y* ont respectivement *k* et ℓ modalités : $x_1, x_2, ..., x_k$ et $y_1, y_2, ..., y_\ell$. Le tableau de contingence décrivant simultanément *X* et *Y* se présente comme suit :



On peut remarquer que: $n_{i\bullet} = \sum_{j=1}^{\ell} n_{ij}$, $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$ et $n_{\bullet\bullet} = \sum_{j=1}^{\ell} n_{\bullet j} = \sum_{i=1}^{k} n_{i\bullet} = \sum_{i=1}^{k}\sum_{j=1}^{\ell} n_{ij} = \sum_{j=1}^{\ell}\sum_{i=1}^{k} n_{ij}$. On rappelle que $f_{j/i} = \frac{n_{ij}}{n_{i\bullet}}$ et $f_{\bullet j} = \frac{n_{\bullet j}}{n}$.

Dans le cas où les variables *X* et *Y* sont indépendantes, $f_{j/i} = f_{\bullet j}, \forall i$ et donc $\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{\bullet j}}{n}$ de telle sorte que $n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$.

Ainsi, les effectifs n_{ij} doivent être égaux **théoriquement** à $\frac{n_{i\bullet} n_{\bullet j}}{n}$ quand X et Y sont indépendantes. On appelle effectifs théoriques que l'on note par T_{ij}, les effectifs définis par $T_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n}$. Il s'agit alors de tester si les effectifs observés, que l'on notera par O_{ij}, s'écarte de manière significative ou non des effectifs théoriques T_{ij}. La démarche du test, que l'on appelle test d'indépendance du khi-deux, est la suivante :

Etape 1 : Détermination des variables X et Y

Etape 2 : Formulation des hypothèses statistiques

H_0 : X et Y sont indépendantes

H_1 : X et Y sont interdépendantes

Etape 3 : Etablissement du seuil de signification α (risque de première espèce)

Etape 4 : Déterminer la valeur critique. La statistique à utiliser est donnée par : $\sum_{j=1}^{r}\left(\frac{(X_j - np_j)^2}{np_j}\right) \sim \chi^2_{r-1}$ en remarquant que dans ce cas $X_j \equiv O_j$, $np_j \equiv T_j$ et *r-1 = (k-1)*$(\ell$*-1)*. Cette façon de calculer le nombre *r-1* de degrés de liberté

s'explique étant donné qu'il y a $k\ell$ données dans le tableau de contingence mais que seulement *(k-1)*$(\ell$*-1)* de ces donnés sont indépendantes du fait de la donnée des deux effectifs partiels.

Etape 5 : Formulation de la règle de décision. Rejeter H_0 si $\chi_c^2 = \sum_{j=1}^{r}\left(\frac{(O_j - T_j)^2}{T_j}\right) > \chi_{r-1,\alpha}^2$, $\chi_{r-1,\alpha}^2$ étant la valeur du khi-deux trouvée dans la tableau à un risque de α.

Etape 6. Décision. Il s'agit de comparer χ_c^2 et $\chi_{r-1,\alpha}^2$ et de décider entre H_0 et H_1.

Remarques.

1. On utilise une simple sommation en considérant que $r = 1,2,...,\ell,\ell+1,...,2\ell,....,k\ell$

2. Dans le cas où les variables *X* et *Y* sont interdépendantes, on peut calculer le coefficient de contingence permettant de mesurer la force d'association entre les deux variables. Ce coefficient est défini par : $C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + n}}$, *0<C<1*. La valeur maximale dépend du nombre de lignes et de colonnes du tableau de contingence. Plus grand est $k\ell$, plus la valeur maximale de *C* se rapproche de 1. Plus grande est la valeur de *C*, plus l'association entre les variables est forte.

3. La distribution d'échantillonnage de χ_c^2 n'est qu'approximativement une distribution khi-deux. De ce fait, il faut être prudent lorsque les effectifs sont petits. On adopte alors comme règle de conduite d'utiliser le test du khi-deux seulement lorsque toutes les fréquences théoriques sont au moins 5. S'il arrivait qu'une ou plusieurs fréquences théoriques soient inferieures à 5, on regroupe alors deux ou plusieurs classes de manière à satisfaire cette condition.

Exemple

Une enquête réalisée dans trois villes des USA a permis d'obtenir les informations suivantes donnant le nombre de consommateurs utilisant favorablement l'un des deux détergents du marché :

O_{ij}

	Détergent A	Détergent B	Total
Los Angeles	242	168	400
San Diego	260	240	500
Fresno	197	203	400
Total	699	611	1300

Peut on accepter que la proportion de consommateurs qui favorisent les deux détergents soit la même ?

T_{ij}

	Détergent A	Détergent B	Total
Los Angeles	215	188	400
San Diego	269	235	500
Fresno	215	188	400
Total	699	611	1300

$T_{11} = \frac{400x699}{1300} = 215$, $T_{12} = \frac{400x611}{1300} = 188$, etc…

2. Test d'ajustement

Quelques méthodes empiriques

La forme de l'histogramme. La forme de l'histogramme construit sur l'échantillon de données peut nous aider à avoir une idée de la distribution de la variable aléatoire dont il est issu. Par exemple, un histogramme symétrique nous orientera par exemple vers une loi normale, de Cauchy, de Student...

La nature du phénomène. Suivant le phénomène étudié, il sera possible d'orienter son choix. Si on s'intéresse à une variable de comptage, on pourra penser à une loi de Poisson, pour une durée de vie on pensera à une loi exponentielle ou à une loi de Weibull... .

Utilisation des moments. On sait que pour une loi de Poisson, la moyenne est égale à la variance. Pour une loi exponentielle la moyenne est égale à l'écart-type. Pour une loi normale le coefficient d'aplatissement (*kurtosis*) est égal à 3 et le coefficient d'asymétrie (*skewness*) est nul.

L'objectif ici est de déterminer si une variable observée se comporte selon une loi de probabilité donnée. Il s'agit alors de tester formellement si une distribution de fréquence observée dans une population ou un échantillon s'ajuste ou obéit à une distribution de probabilité théorique donnée. Si la différence entre ces deux distributions est petite, on conclut que la variable considérée s'ajuste à la distribution de probabilité concernée. La méthode utilisée pour évaluer cette différence s'appelle un test d'ajustement et est tout à fait semblable à celle utilisée dans un test d'indépendance. La différence se situe au niveau des étapes 1 et 2 du test d'indépendance et du nombre de degré de liberté de statistique khi-deux.

Etape 1 : Détermination et définition de la variable d'étude X

Etape 2 : Formulation des hypothèses statistiques

H_0 : $X \sim \mathcal{L}(\theta)$ (*X* suit une distribution de probabilité de paramètre θ)

H_1 : X *ne suit pas* $\mathcal{L}(\theta)$ (*X* ne suit pas une distribution de probabilité de paramètre θ)

θ peut être un scalaire ou vecteur de paramètres.

Etape 3 : Etablissement du seuil de signification α (risque de première espèce)

Etape 4 : Déterminer la valeur critique. La statistique à utiliser est donnée par : $\sum_{j=1}^{r}\left(\frac{(X_j - np_j)^2}{np_j}\right) \sim \chi^2_{r-e-1}$ en remarquant que dans ce cas $X_j \equiv O_j$, $np_j \equiv T_j$, r=nombre de modalités de X et e=nombre de paramètres estimés à partir des observations .

Etape 5 : Formulation de la règle de décision. Rejeter H_0 si $\chi^2_c = \sum_{j=1}^{r}\left(\frac{(O_j - T_j)^2}{T_j}\right) > \chi^2_{r-e-1,\alpha}$, $\chi^2_{r-e-1,\alpha}$ étant la valeur du khi-deux trouvée dans la tableau à un risque de α .

Etape 6. Décision. Il s'agit de comparer χ^2_c et $\chi^2_{r-e-1,\alpha}$ et de décider entre H_0 et H_1.

Remarques.

1. La distribution d'échantillonnage de χ^2_c n'est qu'approximativement une distribution khi-deux. De ce fait, il faut être prudent lorsque les effectifs sont petits. On adopte alors comme règle de conduite d'utiliser le test du khi-deux seulement lorsque toutes les fréquences théoriques sont au moins 5. S'il arrivait qu'une ou plusieurs fréquences théoriques soient inferieures à 5, on regroupe alors deux ou plusieurs classes de manière à satisfaire cette condition.

2. Pour faire le test un test d'ajustement du khi-deux, il est nécessaire de savoir quelle est la loi à tester, c'est-à-dire quelle est sa nature (Normale, Poisson...), mais aussi quels sont ses paramètres. Il est donc souvent nécessaire d'estimer ces paramètres.

Exemple

Sur une population de 200 arbustes, *X* désigne le nombre de fleurs portées par un arbuste. On observé les données suivantes (les modalités 0 à 5 sont regroupées en 5 et les modalités de 13 à 20 sont regroupées en 20, voir remarque précédente):

i	X_i	iX_i	$\hat{p}_i$	$n\hat{p}_i$	$X_i - n\hat{p}_i$	$(X_i - n\hat{p}_i)^2 / n\hat{p}_i$
5	17	85	0.128	25.6	-8.6	2.89
6	19	114	0.096	19.2	-0.2	0.002
7	20	140	0.122	24.4	-4.4	0.79
8	42	336	0.134	26.8	15.2	8.62
9	27	243	0.131	26.2	0.8	0.02
10	25	250	0.115	23.0	2.0	0.17
11	23	253	0.092	18.4	4.6	1.15
12	11	132	0.067	13.4	-2.4	0.43
13	16	208	0.116	23.2	-7.2	2.23
$\sum$	n=200	1761	1.000	-----	-----	16.30

Ces données peuvent elles s'ajuster à une loi de poisson ?

Ces données permettent d'estimer $\hat{\lambda} = 1761/200 = 8.8$ et $\hat{p}_i = \frac{e^{-8.8}(8.8)^i}{i!}$. Attention $\hat{p}_5 = \sum_{i=0}^{5} \hat{p}_i$ et $\hat{p}_{11} = \sum_{i=11}^{20} \hat{p}_i$

2.4.4 Analyse de variance à un facteur – comparaison de plusieurs moyennes

L'analyse de variance à un facteur, permet de déterminer si la différence observée entre plus de deux moyennes est non significative (due au hasard) ou au contraire significative. Par exemple, on peut chercher à vérifier si l'utilisation de différents fertilisants donne des résultats différents en termes de rendement. On parle dans ce cas d'Analyse de Variance (*ANOVA, ANalysis Of VAriance*) à un facteur du fait que l'on cherche à évaluer l'effet d'un seul facteur, les fertilisants, sur le rendement. Si par contre, on cherche à évaluer l'effet du facteur fertilisant par zone géographique, il s'agit d'une ANOVA à deux facteurs par le fait que le rendement observé peut être du et à la qualité du fertilisant utilisé et la localisation géographique des exploitations. Dans le cas qui nous concerne, on s'intéresse à la présentation d'une ANOVA à un facteur.

2.4.4.1 Analyse de variance à un facteur

Pour modéliser le problème d'ANOVA, on prélève n-échantillons indépendants dans *k* populations ; on note X_{ij} la *j-ième* valeur obtenue (j∈{*1, 2, . . ., n*}) dans la population $n^0 i$ (i∈{*1, 2, . . . , k*}). Ces données peuvent être regroupées dans un tableau comme celui-ci :

	1	*2*	*...*	*j*	*...*	*n*	*Moyenne*
1	X_{11}	…	…	…	…	X_{1n}	$\overline{X}_1$
2	…	…	…	…	…	…	…

...	…	…	…	…	…	…	…
i	…	…	…	X_{ij}	…	…	$\overline{X}_i$
...	…	…	…	…	…	…	…
k	X_{k1}	…	…	…	…	X_{kn}	$\overline{X}_k$

On suppose que, pour tout i, les variables aléatoires X_{ij} suivent, pour tout j, la même loi normale $N(\mu_i, \sigma^2)$. On désire tester l'hypothèse nulle : $H_0 : \mu_1 = \mu_2 = ... = \mu_k$. Cette hypothèse peut être formulée de façon alternative si suppose que le modèle générateur des données est décrit par $X_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ où μ représente la moyenne globale, α_i les effets de traitements tels que $\sum \alpha_i = 0$ et ε_{ij} une suite de *nk* variables aléatoires. Ainsi, tester $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ revient à tester $H_0 : \alpha_i = 0$, $\forall i \in \{1, 2, \ldots, k\}$ versus $H_1 : \alpha_i \neq 0$ pour au moins un i.

On rejettera cette hypothèse si la variance des moyennes des échantillons est grande par rapport à la variance dans les échantillons.

Théorème (équation d'analyse de variance) : $\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{\overline{X}})^2 = n\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})^2 + \sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ij} - \overline{X}_i)^2$

S C T = S C E + S C R

Où $\overline{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$, $\overline{\overline{X}} = \frac{1}{k}\sum_{i=1}^{k}\overline{X}_i$ et SCT=Somme Totale des Carrés, SCE= Somme des Carrés expliquée , SCR= Somme des Carrés résiduels.

Démonstration.

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{\overline{X}})^2 &= \sum_{j=1}^{n}\sum_{i=1}^{k}[(\overline{X}_i - \overline{\overline{X}}) + (X_{ij} - \overline{X}_i)]^2 \\ &= \sum_{j=1}^{n}\sum_{i=1}^{k}[(\overline{X}_i - \overline{\overline{X}})^2 + 2(\overline{X}_i - \overline{\overline{X}})(X_{ij} - \overline{X}_i) + (X_{ij} - \overline{X}_i)^2] \\ &= \sum_{j=1}^{n}\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})^2 + 2\sum_{j=1}^{n}\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})(X_{ij} - \overline{X}_i) + \sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ij} - \overline{X}_i)^2 \\ &= n\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})^2 + \sum_{j=1}^{n}\sum_{i=1}^{k}(X_{ij} - \overline{X}_i)^2 \text{ car } \sum_{j=1}^{n}(X_{ij} - \overline{X}_i) = 0.\end{aligned}$$

Sous H_0 et en notant μ la valeur commune des μ_i, la variable aléatoire $\overline{X}_i$ suit, pour tout i, la loi normale $N(\mu, \sigma^2/n)$; donc, d'après le théorème de FISHER, la variable aléatoire $\left(\frac{1}{\sigma^2/n}\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})^2 = \frac{n}{\sigma^2}\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})^2\right) \sim \chi^2_{k-1}$.

Soit $SCE = n\sum_{i=1}^{k}(\overline{X}_i - \overline{\overline{X}})^2$ où SCE = Somme des carrés expliquée (somme des carrés entre échantillons). Ainsi, $\frac{SCE}{\sigma^2} \sim \chi^2_{k-1}$. De même, d'après le théorème de FISHER, les variables aléatoires $\frac{1}{\sigma^2}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2 \sim \chi^2_{n-1}$; elles sont indépendantes. Ainsi, la variable aléatoire $\left(\sum_{i=1}^{k}\left(\frac{1}{\sigma^2}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2\right) = \frac{SCI}{\sigma^2}\right) \sim \chi^2_{k(n-1)}$ en posant

$SCI = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \overline{X}_i)^2$ (somme des carrés à l'intérieur des échantillons). Par conséquent, la statistique $F = \dfrac{\dfrac{SCE}{k-1}}{\dfrac{SCI}{k(n-1)}} \sim F_{(k-1,k(n-1))}$. On rejette H_0 au risque α si $F > F^{\alpha}_{(k-1,k(n-1))}$.

On présente ces résultats sous forme de tableau, dit tableau d'analyse de variance de la façon suivante :

Source de variation de X_{ij}	*Degrés de liberté (ddl)*	*Somme des carrés*	*Moyenne des Sommes des carrés*	*F*
Facteur (traitement)	*k-1*	*SCE*	*SCE/(k-1)*	$F = \dfrac{\dfrac{SCE}{k-1}}{\dfrac{SCI}{k(n-1)}} \sim F_{(k-1,k(n-1))}$
Erreur	*k(n-1)*	*SCR*	*SCR/(kn-k)*	
Total	*kn-1*	*SCT*		

Exemple

On cherche à comparer l'effet de trois détergents différents en réalisant des lectures de blancheur pour un échantillon de 15 morceaux de toile blanche. On met les morceaux de toile dans l'encre de china et on les lave en utilisant une machine à laver avec les trois détergents. On a eu les lectures suivantes :

Détergent A : 77 – 81 – 71 – 76 – 80
Détergent B : 72 – 58 – 74 – 66 – 79
Détergent C : 76 – 85 – 82 – 80 – 77

A un risque de 0.01, les trois détergents sont-ils d'une efficacité identique ?

a) $H_0 : \alpha_i = 0$, $\forall i \in \{1, 2,3\}$ versus $H_1 : \alpha_i \neq 0$ pour au moins un i.

b) $\alpha = 0.01$

c) On rejette H_0 si $F > F^{0.01}_{(2,12)}$

d) On a : T=1125, $T_1 = 385$, $T_2 = 340$, $T_3 = 400$ et $\sum_{i=1}^{3}\sum_{j=1}^{5} X_{ij}^2 = 85041$ et donc SCT=666, SCE=390 et SCR=276.

On présente ces résultats sous forme de tableau, dit tableau d'analyse de variance de la façon suivante :

Source de variation de X_{ij}	*Degrés de liberté (ddl)*	*Somme des carrés*	*Moyenne des Sommes des carrés*	*F*
Facteur (traitement)	*2*	*390*	*195*	

Erreur	*12*	*276*	*23*	$F = \frac{195}{23} = 8.48$
Total	*14*	*666*		

e) On constate que $F = 8.48 > F_{(2,12)}^{0.01} = 6.93$, on rejette l'hypothèse nulle et donc les trois détergents ne sont pas également efficaces.

2.4.4.2 Comparaison multiple post-hoc de moyennes

Une fois que toutes les hypothèses d'une ANOVA ont été vérifiées et que l'analyse a été effectuée, deux conclusions sont possibles, soit on rejette l'hypothèse nulle, soit on n'a pas assez d'évidences pour le faire.

Dans le dernier cas,, généralement' l'analyse s'arrête là. On conclut qu'il n'y a pas de différences significatives entre les groupes. Cependant, dans le premier cas, l'hypothèse d'égalité des groupes est écartée. On veut alors identifier les modalités du résultat significatif. On veut parfois classer les moyennes observées et identifier les différences significatives pour toutes les paires de moyennes.

Pour comparer une paire (*i*, *j*) de moyennes, on formule les hypothèses nulles $H_0^{ij} : \mu_i = \mu_j$. Evidemment, le test global qui compare toutes les moyennes et l'ensemble des tests ''locaux'' qui compare les moyennes deux à deux ne sont pas équivalents. Les hypothèses de base formulées dans le cadre de l'ANOVA sont supposées vérifiées.

Il existe différentes procédures de comparaisons multiples post hoc dans le cadre d'une *ANOVA*. On peut citer par exemple, la procédure *LSD* de Fisher, la procédure *PPDS* de Bonferroni, la procédure *HSD* de Turkey, la procédure de Scheffé, etc…. dans le cas qui nous concerne, on présentera la procédure *LSD* de Fisher et la procédure *PPDS* de Bonferroni.

1. Le ''Least Significant Difference (LSD)'' de Fisher

Hypothèse : $n_i = n_j$

On rejette H_0^{ij} au risque α si la statistique

$$\left| \frac{\overline{Y}_i - \overline{Y}_j}{\sqrt{2\hat{\sigma}^2 / n}} \right| > t_{k(n-1)}^{\alpha/2}$$

où $\hat{\sigma}^2 = \frac{SCR}{k(n-1)}$. En fait, les moyennes $\overline{Y}_i$ et $\overline{Y}_j$ sont déclarées différentes si la valeur absolue de leur différence est supérieure au *LSD* de Fisher où $LSD = t_{k(n-1)}^{\alpha/2} \sqrt{\frac{2\hat{\sigma}^2}{n}}$.

2. Une deuxième version du test de Fisher : le test de Bonferroni

Hypothèse : $n_i \neq n_j$

On rejette H_0^{ij} au risque α^* si la statistique

$$\left| \frac{\overline{Y}_i - \overline{Y}_j}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}} \right| > t_{k(n-1)}^{\alpha^*/2}, \quad \alpha^* = \frac{2\alpha}{k(k-1)}$$

Cette façon de procéder garantit que chaque paire testée, au nombre de $C_k^2 = \frac{k(k-1)}{2}$, s'effectue à un risque α^* qui ne dépasse pas le α global. La démarche de Bonferroni protège le risque global α. En fait, les moyennes $\overline{Y}_i$ et $\overline{Y}_j$ sont déclarées différentes si la valeur absolue de leur différence est supérieure à la ''Plus Petite Difference Significative (*PPDS*)'' où $PPDS = t_{k(n-1)}^{\alpha^*/2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}$, $n = n_i + n_j$

Exercices

Exercice 1. Avant le second tour d'une élection présidentielle, le candidat D. Magog Commande un sondage à une société spécialisée pour savoir s'il a chance d'être élu. Sachant que la proportion p des électeurs qui lui sont favorables ne peut prendre que deux valeurs, ceci conduit au test :

$$\begin{cases} \mathrm{H}_0 : p = 0{,}48 \\ \mathrm{H}_1 : p = 0{,}52 \end{cases}$$

1. Quelle est la signification du choix de p = 0,48 comme hypothèse nulle?
2. Déterminer la région critique du test de risqué α = 0,10 si le sondage a été effectué auprès de n =100 électeurs. Que peut-on penser du résultat ?
3. Indiquer comment varient la région critique et la puissance η en fonction de n. calculer les valeurs de η pour n = 100, 500 et 1 000.
4. On considère maintenant le test :

$$\begin{cases} \mathrm{H}_0 : p = 0{,}49 \\ \mathrm{H}_1 : p = 0{,}51 \end{cases}$$

Calculer la taille d'échantillon minimum n_0 pour que la puissance soit supérieure à 0,90. Quelle est alors sa région critique ?

Exercice 2. Avant d'utiliser un nouvel engrais, un agriculteur décide de l'expérimenter sur cinq parcelles identiques. Les résultats antérieurs permettent de penser que la production X de blé à l'hectare (en quintaux) sur ces parcelles est une *v.a.* de loi $\mathcal{N}(\mathrm{m}, \sigma)$ avec m = 60 et σ = 10. Si on admet que l'utilisation de ce nouvel engrais ne modifie pas la valeur de l'écart type, on doit donc effectuer le test :

$$\begin{cases} \mathrm{H}_0 : m = 60 \\ \mathrm{H}_1 : m = 60 \end{cases}$$

Exercice 3. Le poids de paquets de poudre de lessive, à l'empaquetage, est une v.a. x de loi $\mathcal{N}(m, \sigma)$, d'écart type connu σ = 5g. Le poids marqué sur les paquets est 710g. Toutes les heures, 10 paquets sont tirés au hasard et pesés ; pour une heure donnée on obtient $\overline{\mathrm{X}}_{10}$ =707 g.

1. Tester au risque α = 0,05 l'hypothèse nulle H_0 que le processus d'empaquetage est bien réglé.

2. Déterminer en fonction de m la probabilité d'accepter H_0. Calculer cette valeur pour *m* = 707 g.

Exercice 4. Le stock d'une entreprise est composé de *N* références d'articles, d'une valeur totale réelle inconnue V_R ; le service comptable de l'entreprise a déclaré une valeur V_D. Un vérificateur de la Direction générale des impôts souhaite étudier la validité de la déclaration. Pour cela, il prélève sans remise un échantillon de n références dont il observe les valeurs $(V_1,...,V_n\ )$. On suppose que *n* > 30 et *n/N* < 0.10.

1. Ecrire l'hypothèse nulle H_0 que l'évaluation du stock faite par l'entreprise est correcte, en fonction des éléments fournis. La transformer en faisant intervenir la valeur moyenne réelle inconnue $m_R = V_R / N$ d'une référence et la valeur moyenne $m_D = V_D / N$ calculée à partir de la déclaration fiscale.
2. Donner un estimateur sans biais $\hat{m}_R$ de m_R et sa loi approchée. En déduire la région critique du test de risque α .
3. Application : n =130, N = 1740, α = 0,10, V_D = 130743, moyenne empirique $\bar{v}_n$ = 75,60, écart-type empirique s_n = 18,92. Les données sont exprimées en milliers de gourdes. Conclure pour la validité des comptes de stock.
4. Calculer la probabilité d'accepter l'évaluation du stock alors que la valeur réelle est supérieure de 5% à la valeur déclarée. Même question si la valeur réelle est inférieure de 10% à V_D .

Exercice 5. Une boisson gazeuse, mise en vente au public depuis plusieurs mois, a procuré par quinzaine un chiffre d'affaires de loi normale d'espérance 157000 G et d'écart type 19000G.

Une campagne publicitaire est alors décidée. La moyenne des ventes des huit quinzaines suivant la fin de la promotion est 165000G. On admet que l'écart type reste constant. La campagne publicitaire a-t-elle permis d'accroître le niveau moyen des ventes de 10% ?

Exercice 6. Soit $(X_1,...,X_n)$ un échantillon d'une *v.a.* X de loi normale d'espérance *m* et d'écart- type 1. Déterminer la région critique du test de niveau α_0 =0,05 le plus puissant de :

$$\begin{cases} \mathrm{H}_0 : m \le 3 \\ \mathrm{H}_1 : m > 3 \end{cases}$$

dans le cas où n =100. Déterminer la fonction puissance de ce test.

Exercice 7. La limite du taux X de présent d'un polluant contenu dans des déchets d'usine est 6mg/kg. On effectue un dosage sur 12 prélèvements de 1 kg, pour lesquels on observe les taux $x_{i,}1 \le 12, avec \sum_{i=1}^{12} x_i = 84 et \sum_{i=1}^{12} x_i^2 = 1413.$ On admet que X suit une loi normale $\mathrm{N}(m,\sigma)$.

1. Résoudre le problème de test :
$$\begin{cases} \mathrm{H}_0 : m = 6 \\ \mathrm{H}_1 : m = m_1 (m_1 > 6) \end{cases}$$ avec un risque de première espèce fixé $\alpha = 0.025$.

2. Proposer une région critique pour le test :
$$\begin{cases} \mathrm{H}_0 : m \le 6 \\ \mathrm{H}_1 : m > 6 \end{cases}$$ avec un risque de première espèce fixé $\alpha = 0{,}025.$ L'usine se conforme-t-elle à le législation ?
Peut-on calculer la puissance de ce test ?

Exercice 8. On dispose d'un échantillon de taille n = 15 d'une *v.a*. X de loi normale centrée et d'écart-type $1/\sqrt{\theta}$ où est θ un paramètre inconnu strictement positif.

1. Déterminer la région critique du test de risque α de :
$$\begin{cases} \mathrm{H}_0 : \theta = 1 \\ \mathrm{H}_1 : \theta > 1 \end{cases}$$
Ce test est-il UPP ? Déterminer sa fonction puissance.
2. Quelle décision prend-on pour $\alpha = 0.025$ et $\sum_{i=1}^{15} x_i^2 = 6.8$?
Quelle est la puissance du test pour l'alternative $\mathrm{H}_1 : \theta = 3$?

Exercice 9. Sur le marché, un inspecteur des poids et mesures vérifie la précision de la balance d'un vendeur de fruits et légumes. Il effectue pour cela dix pesées d'un poids étalonné à 100 g et note les indications $x_1,...,x_{10}$ de la balance. On admet que le résultat pesée est une *v.a.* x de loi normale d'espérance 100 et d'écart-type $\sigma = 5$ si la balance est juste et $\sigma > 5$ si elle est déréglée. Ces mesures ayant donné $\sum_{i=1}^{10} (x_i - 100)^2 = 512$, que va en conclure cet inspecteur ?

Exercice 10. L'erreur de mesure d'une grandeur physique est v.a. X de loi $\mathcal{N}(0,2)$, exprimée dans une unité standard. Avec un changement d'unité cette erreur de mesure devient $Y=aX$. Disposant de dix observations, déterminer la région critique du test de risque $\alpha = 0{,}05$ de :
$$\begin{cases} \mathrm{H}_0 : a = 1 \\ \mathrm{H}_1 : a > 1 \end{cases}$$ Ce test est-il UPP ?

Exercice 11. On étudie la résistance à la rupture X d'un fil fabriqué selon des normes ou la résistance moyenne doit être $m_0 = 300g$ avec un écart-type $\sigma_0 = 20g$. La *v.a.* X est supposée suivre une loi normale $\mathcal{N}(\mathrm{m}, \sigma)$. On désire vérifier le respect des normes pour un processus de fabrication nouvellement élaboré.

1. Tester hypothèse sur la base d'un échantillon de 100 bobines de fournissant comme résultats une moyenne empirique $\bar{x}_{100} = 305$ et un écart-type empirique $s_{100} = 22$.
2. Même question pour un échantillon de 10 bobines de moyenne empirique $\bar{x}_{100} = 318$ et d'écart-type empirique $s_{10} = 10$.

Exercice 12. Un généticien veut comparer les proportions p de naissances masculines et 1-p de naissances féminines à l'aide d'un échantillon de n =900 naissances ou on a observé 470 garçons. Il considère donc le test suivant :
$$\begin{cases} \mathrm{H}_0 : p = 0{,}5 \\ \mathrm{H}_1 : p = 0{,}48 \end{cases}$$

1. Quelle est la conclusion sur cet échantillon et pourquoi le généticien est-il peu satisfait de ce test ?
2. Il décide alors d'effectuer le test :
$$\begin{cases} \mathrm{H}_0 : p = 0{,}5 \\ \mathrm{H}_1 : p \neq 0{,}5 \end{cases}$$ Quelle est alors sa conclusion ?

Exercice 13. Une grande administration désirant réduire le nombre horaire de coups de téléphone que ne justifient pas les nécessités du service, envisage d'adopter un dispositif de contrôle dont le fabriquant estime qu'il réduit généralement de moitié nombre de communications. On admet que sans le dispositif le nombre de communications

non justifiées par poste de téléphone et par heure suit une loi de poisson de paramètre $\lambda = 2$. Si le dispositif est efficace, on s'attend à ce que le paramètre de la loi de poisson soit réduit à $\lambda = 1$. On équipe n téléphones avec le dispositif. Soit $x_1,...,x_n$ les nombres de communications observées pour chacun de ces téléphones.

1. Déterminer la région critique du test de risque $\alpha = 0{,}01$ de :

$$\begin{cases} H_0 : \lambda = 2 \\ H_1 : \lambda = 1 \end{cases}$$

Pour n = 6. Calculer le risque de seconde espèce β. Que décide-t-on si $\sum_{i=1}^{6} x_i = 11$

2. On peut vouloir être moins exigeant que le fabricant et tester seulement la diminution du nombre d'appels après mise en service du dispositif. Quelle est la réponse fournie par la méthode de Neyman et Pearson si $\alpha = 0{,}01$ pour le test suivant :

En déduire un test UPP pour :

$$\begin{cases} H_0 : \lambda = 2 \\ H_1 : \lambda < 2 \end{cases}$$ Déterminer l'efficacité du test, représentant $P\left(R^c / H_1\right)$

Exercice 14. Une usine élabore une pâte de verre dont la température de ramollissement X et supposée suivre une loi normale. A six mois d'intervalle, deux séries d'observations sont réalisées et les moyennes et écart-types empiriques sont les suivants : $n_1 = 41, \bar{x}_1 = 785, s_1 = 1{,}68; n_2 = 61, \bar{x}_2 = 788, s_2 = 1{,}40.$ Les deux productions sont-elles identiques ?

Même question avec : $n_1 = 9, \bar{x}_1 = 2510, s_1 = 15{,}9; n_2 = 21, \bar{x}_2 = 2492, s_2 = 24{,}5.$

Exercice 15. Un homme politique s'interrogeant sur ses chances éventuelles de succès aux élections présidentielles commande un sondage qui révèle que, sur 2000 personnes 19% ont l'intention de voter pour lui. Il demande alors à une agence de publicité de promouvoir son image. Un second sondage réalisé après cette campagne publicitaire, auprès de 1000 personnes, montre que 230 d'entre ont l'intention de voter pour lui. Peut-on considérer que cette campagne a été efficace ?

Exercice 16. Dans un atelier de traitements thermiques, on met en service 80 paniers de type I servant à la trempe d'arbres de boites de vitesses et 60 de type II servant à la trempe de pignons de boites de vitesses. Six mois plus tard, il reste en service respectivement 50 paniers sur 80 et 40 sur 60. La résistance à l'usure des deux séries de paniers peut-elle être considérée comme identique ?

Exercice 17. Un laboratoire pharmaceutique désire tester l'efficacité d 'un nouveau médicament. Pour cela il l'applique à un échantillon de 50 malades et compare ces résultats à ceux obtenus avec l'ancien médicament sur 200 malades. Les résultats sont dans le tableau suivant :

	Ancien Médicament	Nouveau médicament
Malades guéris	156	44
Malades non guéris	44	6

Exercice 18. Un programme de génération de nombres au hasard a fourni les valeurs 95, 24, 83,52 et 68. S'agit-il de nombres provenant d'une loi uniforme sur [0,100] ?

Exercice 19. Un examen est ouvert à des étudiants d'origines différentes : économie, informatique et mathématiques. Le responsable de l'examen désire savoir si la formation initiale d'un étudiant influe sur sa réussite. A cette fin, il construit le tableau ci-dessous à partir des résultats obtenus par les 286 candidats, les origines étant précisées en colonne.

	Economie	Informatique	Mathématiques	Total
Réussite	41	59	54	154
Echec	21	36	75	132
Total	62	95	129	286

Quelle est sa conclusion ?

Exercice 20. On étudie la circulation en un point fixe d'une autoroute en comptant, pendant deux heures, le nombre de voitures passant par minute devant un observateur. Le tableau suivant résume les données obtenues :

Nombre de voitures	0	1	2	3	4	5	6	7	8	9	10	11
Fréquence observée	4	9	24	25	22	18	6	5	3	2	1	1

Tester l'adéquation de la loi empirique à une loi théorique simple pour un risque $\alpha = 0{,}10$

Exercice 21. A la sortie d'une chaîne de fabrication, on prélève toutes les trente minutes un lot de 20 pièces mécaniques et on contrôle le nombre de pièces défectueuses du lot. Sur 200 échantillons indépendants on a obtenu les résultats suivants :

Nombre de défectueux	0	1	2	3	4	5	6	7
Nombre de lots	26	52	56	40	20	2	0	4

Tester l'adéquation de la loi empirique du nombre de défectueux par lot de 20 pièces à une loi théorique simple pour un risque $\alpha = 0{,}05$.

Exercice 22. Pour comparer les performances de deux types d'engrais, un cultivateur observe les rendements X_i de n_i parcelles traitées avec l'engrais I $\left(I \le i \le n\right)$ et ceux Y_j de n_2 parcelles traitées avec l'engrais II $\left(1 \le j \le n_2\right)$.
Les espérance respectives sont notées m_1 et m_2.

1. Compte tenu des informations a priori sur la qualité des engrais, on commence par tester :

$$\begin{cases} \mathrm{H}_0 = m_1 = 2, m_2 = 4 \\ \mathrm{H}_1 : m_1 = m_2 = 3 \end{cases}$$

Déterminer par la méthode de *Neyman* et Pearson la région critique de ce test.

2. Déterminer la région critique du test :

$$\begin{cases} \mathrm{H}_0 : m_1 = 2, & m_2 = 4 \\ \mathrm{H}_1 : m_1 \neq 2 & m_2 \neq 4 \end{cases}$$

On utilisera les données numériques suivantes :
$\alpha = 0{,}05, n_1 = 40, n_2 = 60, \sigma = 1, \bar{x}_{40} = 2{,}1$ et $\bar{y}_{60} = 3{,}8$

Exercice 23. Soit X une variable désignant le nombre d'incidents de paiement pour un crédit à la consommation observés sur la durée du prêt. On suppose que X suit une loi de poisson de paramètre $\theta(\theta.0)$. On dispose d'un échantillon de N clients appartenant à une banque A. On note $\{X_1,...,X_N\}$ ce échantillon de ou X_i désigne le nombre d'incidents observés pour l'individu i. On suppose que les variables X_i sont *i.i.d.* de même loi que X et l'on rappelle que :

$$P(X_i = k) = e^{-\theta}\frac{\theta^k}{k!} \quad \text{avec} \quad E(X_i) = \theta \ \text{ et } \ V(X_i) = \theta$$

Exercice 24. On cherche à tester si les clients de cette banque sont en moyenne de ‘’bons’’ clients. Traduisez cette demande sous la forme d'un test d'hypothèse simple contre hypothèse simple, puis sous la forme d'un test d'hypothèse multiple unilatéral.

On considère le test suivant :

$$\mathrm{H}_0 : \theta = \theta_0$$
$$\mathrm{H}_1 : \theta = \theta_1$$

Avec $\theta_0 < \theta_1$. Démontrez que la région critique du test UPP de niveau α est alors de la forme $R = \{X_1,..., X_N / \overline{X}_N > k\}$ où $\overline{X}_N = (1/N)\sum_{i=1}^{N} X_i$ désigne la moyenne empirique des variables X_i.

2. En utilisant le théorème centrale limite, démontrez que dans la cas d'une taille d'échantillon asymptotique $(N \to \infty)$, le seuil critique K associé au test précédent s'écrit sous la forme : $k \approx \theta_0 + \phi^{-1}(1-\alpha)\sqrt{\frac{\theta_0}{N}}$ où $\phi(.)$ désigne la fonction de répartition de la loi normale centrée réduite.

3. On souhaite tester l'hypoth4. On souhaite tester l'hypothèse nulle selon laquelle les clients de le banque A sont faiblement risqués sous la forme suivante : $\mathrm{H}_0 : \theta = 1$ contre $\mathrm{H}_1 : \theta = 2.$ En utilisant les résultats des question 2 et 3, que pouvez vous conclure pour un risque de première espèce de 5% si pour un échantillon de 1332 clients de la banque A on observe les événements suivants

Incidents	0	1	2	3	4	5	6
N_i : Nombre d'individus	510	500	226	70	19	6	1

4. On admet que la région critique du test UPP de niveau $\alpha = 5\%$ de l'hypothèse $\mathrm{H}_1 : \theta = 2$ est définie par : $R = \{X_1,.., X_N / \overline{X}_N > 1.0451\}$. Quel est risque que ce test conduise à déclarer les clients non risqués alors qu'ils sont réellement risqués ?

5. On souhaite tester l'hypothèse nulle selon laquelle les clients de la banque A sont faiblement risqués sous la forme du test unilatéral :

$$\mathrm{H}_0 : \theta = 1$$
$$\mathrm{H}_1 : \theta > 1$$

En utilisant les différents éléments des questions précédentes, que pouvez conclure pour un niveau de risque de 5% ? Vous détaillerez votre démarche.

6. On souhaite enfin tester l'hypothèse :

$$\mathrm{H}_0 : \theta = 1$$
$$\mathrm{H}_1 : \theta \neq 1$$

(i) Construisez la région critique associée à ce test pour un niveau de risque de première espèce de 5% et (ii) concluez à partir des éléments précédents.

7. Donnez la formule de la puissance du test bilatéral (question 7) en fonction de la valeur de θ.

Exercice 25. On considère un échantillon établi par le CEREQ (Centre d'Etudes et de Recherches sur les Qualifications) et constitu130 de 705 jeunes peu diplômés sortis du système scolaire en juin 1989. On s'intéresse à la liaison entre le niveau de salaire des jeunes en euros (noté X) et leur niveau de formation (variable Y). On vous demande de tester au seuil de 5%, puis de 10% l'indépendance du niveau de salaire des jeunes à leur niveau de formation. Vous détaillerez précisément votre démarche.

X / Y	*Bac*	*BEP-CAP*	*Sixième*	*Total*
600 - 750	115	284	30	429
750 - 900	65	109	11	185
900-1650	45	44	2	91
Total	225	437	43	705

Exercice 26. Soit $T = \sum_{i=1}^{k} \sum_{j=1}^{n} X_{ij}$ et $T_i = \sum_{j=1}^{n} X_{ij}$. Démontrer que : a) $SCT = \sum_{i=1}^{k} \sum_{j=1}^{n} X_{ij}^2 - \frac{T^2}{kn}$; b) $T_i = \frac{1}{n} \sum_{i=1}^{k} T_i^2 - \frac{T^2}{kn}$

Exercice 27. On décide de tester l'exactitude des thermostats de trois fers électriques à repasser. On les met à 480^0 F et on observe les températures suivantes :

Fer X : 474 – 496 – 467 – 471
Fer Y : 492 – 498
Fer Z : 460 – 495 - 490

A un risque de 0.01, les températures observées sont elles différentes ?

Exercice 28. Durant cinq semaines consécutives, on a relevé les erreurs commises par quatre techniciens de laboratoire travaillant dans un centre médical :

Technicien 1 : 13 – 16 – 12 - 14 – 15
Technicien 2 : 14 – 16 – 11 - 19 – 15
Technicien 3 : 13 – 18 – 16 – 14 – 18
Technicien 4 : 18 – 10 – 14 – 15 – 12

A un risque de 0.05, la différence entre le nombre d'erreurs commises peut elle être due au hasard ?