Capes de Mathématiques
Petit cours de Statistique inférentielle

Jeanne Fine

Septembre 2007

Table des matières

1 Introduction

1.1 Échantillonnage, estimation et tests

Lors d'une étude statistique, il est rare que l'on puisse obtenir l'information auprès de toute la population. On réalise alors l'étude sur un échantillon dit aléatoire ou probabiliste, c'est-à-dire, obtenu par une procédure aléatoire contrôlée ou supposé obtenu ainsi.

L'objectif est alors d'inférer les résultats obtenus sur l'échantillon à la population tout entière. Cette procédure sera réalisée en deux étapes. Les caractéristiques (moyenne, écart-type, ...) calculées sur l'échantillon varient d'un échantillon à l'autre, ce sont des observations de variables aléatoires, dites d'échantillonnage. La première étape (déductive), appelée *théorie de l'échantillonnage*, consiste à déterminer (grâce au calcul des probabilités) les distributions de probabilité (au moins approchées) de ces variables d'échantillonnage. La deuxième étape (inductive), appelée *théorie de l'estimation*, consiste, à partir des caractéristiques calculées sur **1** échantillon, à estimer les caractéristiques inconnues (les *paramètres*) de la population. Les résultats obtenus lors de la première étape permettent, dans la deuxième, de fournir un encadrement (*intervalle de confiance*) du paramètre à estimer avec une certaine probabilité (fixée à l'avance) de se tromper.

Enfin une troisième partie de la statistique inférentielle, appelée *théorie des tests*, consiste à établir des règles de décision permettant de conclure si une différence observée - entre le résultat calculé sur l'échantillon et le résultat attendu sur la population ou bien entre les résultats calculés sur deux échantillons - provient de la *fluctuation d'échantillonnage* (puisque ce sont des observations de variables aléatoires) ou bien si la différence est *significative*. Elle révèle alors une différence qui n'est pas due au hasard. Ces procédures de décision sont toujours accompagnées d'une probabilité de se tromper fixée à l'avance.

L'approche classique de la statistique inférentielle repose sur la donnée d'un échantillon $(X_i)_{i=1,\dots,n}$ de n variables aléatoires réelles indépendantes et identiquement distribuées comme X, c'est-à-dire, de même distribution de probabilité que X, ce que l'on notera n v.a.r. i.i.d. comme X. On suppose que la v.a.r. X est définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathcal{P})$ associé à une expérience aléatoire $\mathcal{E}$ et qu'elle admet une espérance mathématique μ et une variance σ^2. L'échantillon de v.a.r. est alors obtenu à partir de la répétition n fois de suite et dans les mêmes conditions de l'expérience $\mathcal{E}$.

1.2 Approche sondage

Afin de mieux comprendre et interpréter les notions introduites en statistique inférentielle, on introduit "l'approche sondage" : on tire "au hasard" (c'est-à-dire, avec équiprobabilité) un individu statistique ω dans une population finie Ω de taille N. Le modèle probabilisé associé à cette expérience aléatoire est donc $(\Omega, \mathcal{P}(\Omega), P)$ où P est l'équiprobabilité. On s'intéresse à l'observation sur cet individu d'un caractère quantitatif $\mathcal{X}$: $x = \mathcal{X}(\omega)$. Alors x est l'observation d'une v.a.r. X définie sur $(\Omega, \mathcal{P}(\Omega), P)$ dont la loi de probabilité P_X n'est autre que la distribution de fréquence de $\mathcal{X}$ sur la population Ω.

On a en effet : $P_X(\{x\}) = P([X = x]) = P(\{\omega \in \Omega\,;\, \mathcal{X}(\omega) = x\}) = \frac{N_x}{N}$, où N_x est le

nombre d'individus de la population Ω prenant la valeur x pour le caractère $\mathcal{X}$.

On en déduit :

$$\mu = \frac{1}{N}\sum_{\omega\in\Omega} X(\omega) = \sum_{x\in X(\Omega)} xP([X=x]) = \mathbb{E}(X) \text{ et}$$

$$\sigma^2 = \frac{1}{N}\sum_{\omega\in\Omega}(X(\omega))^2 - \mu^2 = \sum_{x\in X(\Omega)} x^2P([X=x]) - \mu^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

La v.a.r. X est donc définie sur $(\Omega, \mathcal{P}(\Omega), P)$ et à valeurs dans $(\Omega_X, \mathcal{P}(\Omega_X), P_X)$ avec $\Omega_X = X(\Omega)$. Dans le cas particulier où $\mathcal{X}$ est l'indicatrice d'une sous-population A de Ω, alors on a : $\mu = p$ avec $p = \text{card}(A)/N$, proportion d'individus de Ω appartenant à A, et $\sigma^2 = p(1-p)$ (car $\mathcal{X}$ ne prenant que les valeurs 0 ou 1 vérifie $\mathcal{X}^2 = \mathcal{X}$). La v.a.r. X suit alors une loi de Bernoulli de paramètre p.

Un échantillon de taille n i.i.d. comme X, $(X_i)_{i=1,...,n}$, est obtenu à partir du *tirage avec remise* d'un échantillon de taille n dans Ω. L'ensemble Ω' des échantillons d'individus avec remise de taille n est alors $\Omega' = \Omega^n$ (d'où $\text{card}(\Omega') = N^n$) et tous ces échantillons ont la même probabilité de sortir. Le tirage aléatoire avec remise d'un échantillon d'individus de taille n est donc modélisé par $(\Omega', \mathcal{P}(\Omega'), P')$ avec $\Omega' = \Omega^n$ et P' l'équiprobabilité sur $(\Omega', \mathcal{P}(\Omega')$. On peut remarquer que l'on a :

$$\forall (A_1, ..., A_n) \subset \Omega^n,\ P'(\prod_{i=1}^{n} A_i) = \frac{\text{card}(\prod_{i=1}^{n} \text{A}_\text{i})}{N^n} = \prod_{i=1}^{n}(\frac{\text{card}(\text{A}_\text{i})}{N}) = \prod_{i=1}^{n} P(A_i).$$

La probabilité P' traduit l'indépendance des événements associés aux différents tirages, c'est la *probabilité produit* n fois de P et elle est notée $P^{\otimes n}$.

Le vecteur aléatoire $X' = (X_1, ..., X_n)$ est donc défini sur $(\Omega', \mathcal{P}(\Omega'), P')$ et à valeurs dans $(\Omega'_X, \mathcal{P}(\Omega'_X))$; la loi de probabilité de X' (image de P' par X') n'est autre que la loi de probabilité produit n fois de P_X, notée $(P_X)^{\otimes n}$, qui traduit l'indépendance des v.a.r. $(X_i)_i = 1, ..., n$:

$$\forall x' = (x_1, ..., x_n) \in \Omega'^n_X,\ P'([X'=x'] = P'(\bigcap_{i=1}^{n}[X_i = x_i]) = \prod_{i=1}^{n} P'([X_i = x_i]) = \prod_{i=1}^{n} P_X(\{x_i\}).$$

On peut résumer ce modèle d'échantillonnage i.i.d. dans le diagramme commutatif suivant, commutatif car on a : $Q = (P')_{X'} = (P_X)^{\otimes n}$.

$$\begin{array}{ccc}
\textbf{1}\text{ expérience aléatoire } \mathcal{E} & & \\
(\Omega, \mathcal{P}(\Omega), P) & \xrightarrow{X} & (\Omega_X, \mathcal{P}(\Omega_X), P_X) \\
\Downarrow & & \Downarrow \\
n \text{ exp. aléatoires ind.} & & \\
(\Omega^n, \mathcal{P}(\Omega^n), P') & \xrightarrow{X'=(X_1,...,X_n)} & (\Omega'^n_X, \mathcal{P}(\Omega'^n_X), Q) \\
\text{avec } P' = P^{\otimes n} & & \text{avec } Q = (P')_{X'} = (P_X)^{\otimes n}
\end{array}$$

Le vecteur aléatoire X' associe à $\omega' = (\omega_1, ..., \omega_n)$, *échantillon d'individus de taille n* tiré de Ω à probabilité égale et avec remise, un vecteur $x' = (x_1, ..., x_n)$, *échantillon de n valeurs* de X, observations d'un *échantillon de n v.a.r.* $X' = (X_1, ..., X_n)$ i.i.d. comme X.

Remarque

Lorsque l'on tire de Ω un échantillon de taille n à probabilité égale mais *sans* remise, alors Ω' est l'ensemble des parties de n éléments de Ω (d'où $\text{card}(\Omega') = \binom{N}{n}$) et tous les échantillons ont la même probabilité de sortir, donc P' est l'équiprobabilité sur $(\Omega', \mathcal{P}(\Omega'))$. Les v.a.r. $(X_i)_{i=1,...,n}$, où X_i est la v.a.r. donnant la valeur de X observée sur le $i^{ème}$ individu de l'échantillon, sont encore identiquement distribuées comme X mais ne sont plus indépendantes. Dans la pratique, lorsque le taux de sondage n/N est faible (≤ 0.10) on assimile un échantillon sans remise à un échantillon avec remise.

1.3 Approche classique

Comme déjà dit en introduction, dans cette approche, on se donne une v.a.r. X définie sur un espace probabilisé $(\Omega, \mathcal{A}, P)$ associé à une expérience aléatoire mais on ne fait plus référence à l'espace de départ. On ne s'intéresse donc qu'à la partie droite du diagramme commutatif du § 1.2.

On envisage quatre cas pour la loi de probabilité P_X de la v.a.r. X, un cas général et trois cas particuliers. Dans le cas général, on ne précise pas cette loi de probabilité, on sait seulement que X admet une moyenne μ et une variance σ^2 (donc un écart-type σ), ce que l'on notera : $X \sim P_X(\mu, \sigma)$.

1. Cas général : $X \sim P_X(\mu, \sigma)$
2. Cas fini : X finie de loi de probabilité $\{(x^h, p^h); h = 1, ..., H\}$; c'est un cas particulier du 1) et on a : $\mu = \sum_{h=1}^{H} x^h p^h$ et $\sigma^2 = \sum_{h=1}^{H} (x^h)^2 p^h - \mu^2$
3. Cas normal : $X \sim N(\mu, \sigma)$
4. Cas Bernoulli : $X \sim B(1, p)$; on a alors : $\mu = p$ et $\sigma^2 = p(1-p)$.

On remarquera que l'approche sondage d'un échantillon avec remise (et même d'un échantillon sans remise dans le cas où le taux de sondage n/N est négligeable) rentre dans le cas général, et même dans le cas fini avec $p^h = N^h/N$ où N^h est le nombre d'individus de la population qui prennent la valeur x^h de X.

2 Échantillonnage

2.1 Lois du Khi2 et de Student

Si $(X_i)_{i=1,...,n}$ sont des v.a.r. indépendantes de loi normale centrée réduite $N(0\,;1)$, alors, par définition, $\Sigma_{i=1}^{n} X_i^2$ suit une loi de $Khi2$ à n degrés de liberté, notée $\chi^2_{(n)}$.

Si X et Y sont deux v.a.r. indépendantes, $X \sim N(0\,;1)$ et $Y \sim \chi^2_n$, alors, par définition, $X/\sqrt{Y/n}$ suit une loi de *Student* à n degrés de liberté, notée $St_{(n)}$.

Pour n assez grand (dans la pratique, $n \geq 30$), on a approximativement : $St_{(n)} \sim N(0\,;1)$.

2.2 Variables d'échantillonnage, espérance mathématique et variance

Variables d'échantillonnage

Soit $(X_i)_{i=1,...,n}$ un échantillon de n v.a.r. i.i.d. comme X. On pose :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i,\ V_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 \text{ et } S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

Ces trois v.a.r. sont respectivement appelées *moyenne d'échantillonnage, variance d'échantillonnage* et *variance corrigée d'échantillonnage.*

Dans le cas 4) où X est une v.a.r. de Bernoulli de paramètre p (p étant la probabilité de succès lors d'une expérience) $\overline{X}_n$ est la fréquence de succès lors de n expériences. Cette v.a.r. $\overline{X}_n$ est alors notée F_n et appelée *fréquence d'échantillonnage.*
On a alors $V_n = F_n(1 - F_n)$ donc $S_n^2 = \frac{n}{n-1}F_n(1 - F_n)$. $(*)$

Dans le cas 2) d'une v.a.r. X finie, si $(x_i)_{i=1,...,n}$ est une observation de l'échantillon de v.a.r. $(X_i)_{i=1,...,n}$, la distribution de fréquence est notée $\{(x^h, f_n^h); h = 1, ..., H\}$ où f_n^h est la fréquence de valeurs x_i égales à x^h, alors $(f_n^1, ..., f_n^H)$ est l'observation d'un vecteur aléatoire $(F_n^1, ..., F_n^H)$ (implicitement associée au vecteur $(x^1, ..., x^H)$) appelé *distribution de fréquence d'échantillonnage.*

Espérance mathématique et variance des variables d'échantillonnage

Il est aisé de vérifier, pour tout $n \geq 2$:

$$\mathbb{E}(\overline{X}_n) = \mu,\ \text{var}(\overline{X}_n) = \frac{\sigma^2}{n} \text{ et } \mathbb{E}(V_n) = \frac{n-1}{n}\sigma^2$$

On en déduit, d'une part : $\mathbb{E}(S_n^2) = \sigma^2$ et $\mathbb{E}(\frac{S_n^2}{n}) = \text{var}(\overline{X}_n) = \frac{\sigma^2}{n}$, $(**)$

d'autre part : $\mathbb{E}(F_n) = p,\ \text{var}(F_n) = \frac{p(1-p)}{n}$ et $\mathbb{E}(\frac{F_n(1-F_n)}{n-1}) = \text{var}(F_n)$ d'après $(*)$ et $(**)$.

Si on pose : $U_n = (\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}})$ et, dans le cas 4) : $U_n = (\frac{F_n - p}{\sqrt{p(1-p)/n}})$
alors U_n est une v.a.r. centrée réduite.

2.3 Distributions de probabilité des variables d'échantillonnage

2.3.1 Loi des grands nombres et "loi vulgarisée" des grands nombres

Loi des grands nombres

La suite de v.a.r. $(\overline{X}_n)$ converge en probabilité vers μ lorsque la taille n de l'échantillon converge vers l'infini ; la v.a.r. $\overline{X}_n$ est asymptotiquement une "variable constante" égale à μ.

En effet, pour tout $n \geq 2$, $\mathbb{E}(\overline{X}_n) = \mu$ et $\text{var}(\overline{X}_n) = \frac{\sigma^2}{n}$, donc $\lim\limits_{n\to\infty} \text{var}(\overline{X}_n) = 0$ et l'on en déduit le résultat.

De même, dans le cas 4), la v.a.r. F_n est asymptotiquement une "variable constante" égale à p. Il s'agit de "l'approche fréquentiste" des probabilités : la fréquence de succès lors de n expériences aléatoires indépendantes converge vers la probabilité de succès lors d'une expérience.

"Loi vulgarisée des grands nombres"

Lorsque $(F_n^1, ..., F_n^H)$ est la distribution de fréquence d'échantillonnage d'une v.a. catégorielle ou d'une v.a.r. finie (cas 2) dont la distribution de probabilité est $(p^1, ..., p^H)$, la distribution de fréquence $(F_n^1, ..., F_n^H)$ est asymptotiquement égale à la distribution de probabilité $(p^1, ..., p^H)$. C'est ainsi qu'est énoncée la "loi vulgarisée des grands nombres" dans les programmes de terminale des lycées.

Dans le cas 2) d'une v.a.r. X finie, si $(x_1, ..., x_n)$ est l'observation d'un échantillon de n v.a.r. $(X_1, ..., X_n)$ i.i.d. comme X, alors $\overline{x}_n = \frac{1}{n}\sum_{i=1}^n x_i$ est l'observation de la moyenne d'échantillonnage : $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Comme on a aussi : $\overline{x}_n = \sum_{h=1}^H x^h f_n^h$ et $\overline{X}_n = \sum_{h=1}^H x^h F_n^h$, la loi vulgarisée des grands nombres permet de retrouver le fait que la v.a.r. $\overline{X}_n = \sum_{h=1}^H x^h F_n^h$ est asymptotiquement égale à : $\sum_{h=1}^H x^h p^h$ c'est-à-dire μ.

Simulation

La loi vulgarisée des grands nombres permet d'approcher la distribution de probabilité inconnue d'une v.a.r. à partir des valeurs $(x_i)_{i=1,...K}$ (avec, par exemple, $K = 100$) d'un échantillon de v.a.r. $(X_i)_{i=1,...,K}$ i.i.d. comme X. L'objet de la *simulation* est de produire de tels échantillons.

Lorsque l'objectif est d'approcher par simulation la distribution de probabilité de la v.a.r. $\overline{X}_n$, moyenne d'un échantillon de v.a.r. $(X_1, ..., X_n)$ i.i.d. comme X, on aura à simuler K échantillons de taille n de X, donc $K \times n$ valeurs de X.

2.3.2 Théorème central limite et distributions de probabilité approchées des variables d'échantillonnage

On rappelle que l'on a : $U_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ avec $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$.

Théorème central limite

La suite de v.a.r. centrées réduites (U_n) converge en loi vers une v.a.r. normale centrée réduite lorsque la taille de l'échantillon n tend vers l'infini.

Approximations

Pour n fixé suffisamment grand, $(n \geq 30)$, la distribution de probabilité approchée de $\overline{X}_n$ est donnée par :

$$\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}\right) \overset{approx}{\sim} N(0\,;1)$$

On admettra que le théorème reste valable lorsque l'on remplace σ par la v.a.r. S_n (dont l'espérance mathématique du carré est égale à σ^2).

On a donc aussi, pour n fixé suffisamment grand, $(n \geq 30)$:

$$(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}) \overset{approx}{\sim} N(0\,;1)$$

Dans le cas 3) où X est normale, on a pour tout $n \geq 2$,

$$(\frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}) \sim N(0;1) \quad \text{et} \quad \sum_{i=1}^{n}(\frac{X_i - \mu}{\sigma})^2 \sim \mathcal{X}^2_{(n)}$$

On peut écrire :

$$\sum_{i=1}^{n}(\frac{X_i - \overline{X}_n}{\sigma})^2 = (n-1)\frac{S_n^2}{\sigma^2}$$

et on admettra les deux résultats suivants :

$(n-1)\frac{S_n^2}{\sigma^2} \sim \mathcal{X}^2_{(n-1)}$ et les deux v.a.r. $\overline{X}_n$ et S_n^2 sont indépendantes en probabilité.

On en déduit, pour tout $n \geq 2$:

$$(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}) \sim St_{(n-1)}$$

et on retrouve le résultat précédent pour n suffisamment grand, $(n \geq 30)$:

$$(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}}) \overset{approx}{\sim} N(0\,;1).$$

Dans le cas 4) où X est une variable de Bernoulli, on a, pour n fixé suffisamment grand, $(n \geq 30)$:

$$(\frac{F_n - p}{\sqrt{p(1-p)/n}}) \overset{approx}{\sim} N(0;1) \quad \text{et} \quad (\frac{F_n - p}{\sqrt{F_n(1-F_n)/(n-1)}}) \overset{approx}{\sim} N(0\,;1)$$

Lorsque $(F_n^1, ..., F_n^H)$ est la distribution de fréquence d'échantillonnage d'une v.a. catégorielle ou d'une v.a.r. finie (cas 2) dont la distribution de probabilité est $(p^1, ..., p^H)$, pour n suffisamment grand $(n \geq 30)$ on admettra :

$$n\sum_{h=1}^{H}\frac{(F_n^h - p^h)^2}{p^h} \overset{approx}{\sim} \chi^2_{(H-1)}.$$

On utilisera cette propriété au § 4.3 pour le test du Khi2 d'adéquation d'une distribution de fréquence à une distribution de probabilité.

3 Estimation

3.1 Estimation ponctuelle

Comme l'espérance mathématique de $\overline{X}_n$ est égale à μ, à partir d'un échantillon de taille n, on estimera ponctuellement μ par $\overline{X}_n$. On dit que la v.a.r. $\overline{X}_n$ est un estimateur sans biais de μ (cf. § 3.4).

De même, la v.a.r. S_n^2 est un estimateur sans biais de σ^2, aussi, on estimera ponctuellement σ par l'estimateur S_n (estimateur biaisé car $E(S_n) \neq \sqrt{E(S_n^2)}$). La v.a.r. S_n^2/n est un estimateur sans biais de var$(\overline{X}_n)$.

Dans le cas 4), la v.a.r. F_n est un estimateur sans biais de p et la v.a.r. $F_n(1-F_n)/(n-1)$ est un estimateur sans biais de var(F_n). Néanmoins, on estime généralement l'écart-type de F_n par $\sqrt{F_n(1-F_n)/n}$.

3.2 Estimation par intervalle de confiance d'une moyenne (et d'une proportion)

"Moyenne" et "proportion" sont le vocabulaire utilisé dans l'approche sondage. Dans le cas général, on devrait parler d'"espérance mathématique" et de "probabilité". Cette remarque est valable pour toute la suite du cours.

Soit $\alpha \in]0,1[$, on note u_α (resp. $t_{n-1,\alpha}$) le *quantile* d'ordre α de la v.a.r. U de loi $N(0,1)$ (resp. de la v.a.r. T_{n-1} de loi $St_{(n-1)}$), c'est-à-dire, le réel vérifiant $P(U \leq u_\alpha) = \alpha$ (resp. $P(T_{n-1} \leq t_{n-1,\alpha}) = \alpha$). Ces deux lois étant symétriques, on en déduit :

$$P(-u_{1-\frac{\alpha}{2}} \leq U \leq u_{1-\frac{\alpha}{2}}) = 1-\alpha \text{ et } P(-t_{n-1,1-\frac{\alpha}{2}} \leq T_{n-1} \leq t_{n-1,1-\frac{\alpha}{2}}) = 1-\alpha.$$

Cas général

Dans le cas général 1) et dans le cas où σ est inconnu et estimé par S_n, on a, pour n suffisamment grand :

$$P(-u_{1-\frac{\alpha}{2}} \leq \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leq u_{1-\frac{\alpha}{2}}) = 1-\alpha, \text{ ou encore :}$$

$$P(\overline{X}_n - u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \leq \mu \leq \overline{X}_n + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}) = 1-\alpha.$$

L'interprétation de cette dernière écriture est la suivante : la probabilité que l'intervalle aléatoire $[\overline{X}_n - u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \ ; \ \overline{X}_n + u_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}]$ contienne le paramètre inconnu μ est égale à $1-\alpha$.

A partir d'**1** échantillon, on déduit du résultat précédent *l'estimation de μ par intervalle de confiance, au niveau de confiance* 1 - α :

$$[\overline{x}_n - u_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}} \ ; \ \overline{x}_n + u_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}]$$

On a remplacé les v.a.r. par leurs observations sur l'échantillon dont on dispose.

Cas particuliers

Bien sûr dans le cas où σ est connu, il n'est pas utile de l'estimer et l'estimation de μ par intervalle de confiance au niveau $1-\alpha$ est alors :

$$[\overline{x}_n - u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \ ; \ \overline{x}_n + u_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}]$$

Dans le cas normal 3) lorsque σ est connu, cet intervalle de confiance peut être construit quelque soit la taille n de l'échantillon (supérieure à 2).

Dans le cas normal 3) lorsque σ est inconnu et estimé par S_n, on construit alors l'intervalle de confiance en utilisant la loi de Student à $(n-1)$ degrés de liberté :

$$[\overline{x}_n - t_{n-1,1-\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}} \; ; \; \overline{x}_n + t_{n-1,1-\frac{\alpha}{2}} \frac{s_n}{\sqrt{n}}]$$

Dès que la taille n de l'échantillon est supérieure à 30, on peut réutiliser la loi normale centrée réduite.

Dans le cas 4) de Bernoulli, on a, pour n suffisamment grand, l'estimation de p par intervalle de confiance au niveau $1-\alpha$ suivant :

$$[f_n - u_{1-\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}} \; ; \; f_n - u_{1-\frac{\alpha}{2}} \sqrt{\frac{f_n(1-f_n)}{n}}].$$

Comme, pour tout x de $[0\,,1]$, on a $x(1-x) \leq 1/4$, l'intervalle de confiance de p à 95% de confiance ($\alpha = 5\%$ et $u_{1-\frac{\alpha}{2}} = 1.96$), est contenu dans le suivant :

$$[f_n - \frac{1}{\sqrt{n}} \; ; f_n + \frac{1}{\sqrt{n}}].$$

On en déduit que pour $n = 100$, l'intervalle est $f_n \pm 10\%$, pour $n = 1000$, l'intervalle est $f_n \pm 3\%$, etc. (Cf. la fiche n° 9 des "Onze fiches de statistique" du document d'accompagnement des programmes de la classe de seconde" et le document d'accompagnement des programmes des classes de terminales : sondages.)

Commentaire et vocabulaire

a) *Confiance et probabilité*

On lit dans certains programmes de BTS de Lycées Techniques :
"On distinguera confiance et probabilité :
- avant le tirage d'un échantillon, la procédure d'obtention de l'intervalle de confiance a une probabilité $1-\alpha$ que cet intervalle contienne le paramètre inconnu,
- après le tirage, le paramètre est dans l'intervalle calculé avec une confiance de $1-\alpha$."

Supposons α=5%. Si l'on une confiance à 95%, c'est que l'on peut se tromper dans 5% des cas. Dans l'approche sondage, on peut dire que 95% des échantillons aléatoires simples avec remise de taille n produiront un intervalle (construit comme ci-dessus) contenant μ. On espère que l'échantillon dont on dispose fait partie de ces 95%.

Dans l'approche classique, le nombre d'échantillons i.i.d. comme X de taille n peut être infini. On a cependant une probabilité égale à 0.95 qu'un échantillon i.i.d. comme X de taille n fournisse un intervalle (construit comme ci-dessus) contenant μ.

Si on simule un grand nombre K d'échantillons de taille n i.i.d. comme X, alors environ 95% des échantillons simulés fourniront un intervalle (construit comme ci-dessus) contenant μ.

b) *Intervalle de dispersion (ou de probabilité) et intervalle de confiance*

Apparaît dans les ouvrages de vulgarisation, le vocabulaire "intervalle de dispersion à 95%" ou "intervalle de probabilité de 95%", qui correspond pour la v.a.r. $\overline{X}_n$, à l'intervalle

$$[\mu - 1.96\frac{s_n}{\sqrt{n}} \; ; \; \mu + 1.96\frac{s_n}{\sqrt{n}}].$$

La probabilité que la v.a.r. $\overline{X}_n$ appartienne à l'intervalle est 0.95. En revanche, lorsque l'on remplace la v.a.r. $\overline{X}_n$ par son observation sur l'échantillon dont on dispose, on construit une estimation de μ par intervalle de confiance à 95%

$$[\overline{x}_n - 1.96\frac{s_n}{\sqrt{n}} \; ; \; \overline{x}_n + 1.96\frac{s_n}{\sqrt{n}}].$$

c) *Intervalle de confiance et fourchette de sondage*

Fourchette de sondage est le terme utilisé par les médias dans le cadre des sondages d'opinion pour désigner un intervalle de confiance. Il s'agit de l'intervalle de confiance pour l'estimation d'une proportion présenté ci-dessus (cas particulier 4 de Bernoulli). La "probabilité" à estimer est dans ce cas la proportion, notée p, de la sous-population d'intérêt dans la population totale. Elle est estimée ponctuellement par la proportion f_n de cette sous-population dans l'échantillon.

3.3 Exercice : estimation d'une différence de deux moyennes

3.3.1 Échantillons indépendants (une variable, deux populations)

Soit X une variable réelle définie sur une population Ω ; Ω_1 et Ω_2 deux sous-populations disjointes.

Soit μ_1 et σ_1 la moyenne et l'écart-type de X sur Ω_1, μ_2 et σ_2 la moyenne et l'écart-type de X sur Ω_2.

On tire, de façon indépendante, deux échantillons aléatoires simples à probabilités égales avec remise de taille n_1 et n_2 dans les sous-populations Ω_1 et Ω_2. Il s'agit d'un échantillon aléatoire dit stratifié de taille $n_1 + n_2$.

Soit $\overline{X}_{n1}$ et S^2_{n1}, $\overline{X}_{n2}$ et S^2_{n2} les moyennes et variances (corrigées) d'échantillonnage correspondant aux deux échantillons.

On pose $\delta = \mu_1 - \mu_2$ et $\widehat{\delta} = \overline{X}_{n1} - \overline{X}_{n2}$.

Montrer que $\widehat{\delta}$ est un estimateur sans biais de δ.

Calculer la variance de $\widehat{\delta}$ et donner un estimateur sans biais de cette variance à l'aide des variables d'échantillonnage.

En admettant l'approximation normale pour $\widehat{\delta}$, déterminer un intervalle de confiance à 95% de δ.

3.3.2 Échantillons appariés (une population, deux variables)

Soit X et Y deux variables réelles définies sur une population Ω.

Soit μ_X et σ_X , μ_Y et σ_Y la moyenne et l'écart-type de X et de Y sur la population.

On tire un échantillon aléatoire simple à probabilités égales avec remise de taille n.

Soit $\overline{X}_n$ et S^2_{nX}, $\overline{Y}_n$ et S^2_{nY} la moyenne et variance d'échantillonnage de X et de Y.

On pose $\delta = \mu_X - \mu_Y$ et $\widehat{\delta} = \overline{X}_n - \overline{Y}_n$.

Montrer que $\widehat{\delta}$ est un estimateur sans biais de δ.

On introduit la variable $D = X - Y$ et les variables d'échantillonnage $\overline{D}_n$ et S^2_{nD}.

Calculer la variance de $\widehat{\delta}$ et donner un estimateur sans biais de cette variance à l'aide des variables d'échantillonnage.

En admettant l'approximation normale pour $\widehat{\delta}$, déterminer un intervalle de confiance à 95% de δ.

Remarque

On a utilisé l'approche sondage pour la présentation de ces deux types d'échantillon (*échantillons indépendants* et *échantillons appariés*).

Dans l'approche classique, on a, dans le premier cas, deux échantillons indépendants entre eux de v.a.r. $(X_{i1})_{i=1,\ldots,n_1}$ i.i.d. comme une v.a.r. X_1 et de v.a.r. $(X_{i2})_{i=1,\ldots,n_2}$ i.i.d. comme une v.a.r. X_2, dans le deuxième cas, un échantillon de couples de v.a.r. $(X_i, Y_i)_{i=1,\ldots,n}$ i.i.d. comme un couple de v.a.r. (X, Y).

Dans les deux cas, on peut se poser des questions sur l'égalité des moyennes ou des variances ou des distributions de probabilité des deux v.a.r. sous-jacentes.

3.4 Compléments théoriques sur l'estimation

3.4.1 Estimateurs et estimations

On considère un échantillon statistique $(X_i)_{i=1,\ldots,n}$ i.i.d. comme X, v.a.r. d'ordre 2. La loi de probabilité de X dépend d'un paramètre réel θ. Par exemple, dans les cas 1) à 3) $\theta = \mu$ ou $\theta = \sigma$, dans le cas 4) $\theta = p$.

Définitions

Un *estimateur* $\widehat{\theta}_n$ de θ est une v.a.r. fonction des v.a.r. $(X_i)_{i=1,\ldots,n}$ ayant de "bonnes propriétés" par rapport à θ (cf. § suivant) ; une *estimation* est une observation de cette v.a.r. (estimation ponctuelle).

L'estimation (ou *théorie de l'estimation*) consiste à étudier les critères d'estimation et les propriétés des estimateurs obtenus.

3.4.2 Critères de qualité d'un estimateur

Soit θ_n un estimateur de θ.

Définitions

1) *le biais* de θ_n, $B(\theta_n)$, et défini par : $B(\theta_n) = \mathbb{E}(\theta_n) - \theta$

2) *l'erreur quadratique moyenne* de θ_n, $EQM(\theta_n)$, est définie par : $EQM(\theta_n) = \mathbb{E}[(\theta_n - \theta)^2]$ et elle vérifie : $EQM(\theta_n) = \text{var}(\theta_n) + [B(\theta_n)]^2$.

3) θ_n est un *estimateur sans biais* de θ si, et seulement si, son biais est nul, c'est-à-dire : $\mathbb{E}(\theta_n) = \theta$

4) θ_n est un *estimateur convergent* de θ si, et seulement si, $(\theta_n) \xrightarrow[n\to\infty]{\text{Pr}} \theta$.

On rappelle la propriété déjà utilisée précédemment :

$$(\lim_{n\to\infty} \mathbb{E}(\theta_n) = \theta \text{ et } \lim_{n\to\infty} \text{var}(\theta_n) = 0) \Longrightarrow (\theta_n) \xrightarrow[n\to\infty]{\text{Pr}} \theta$$

Remarques

Un estimateur est d'autant plus précis que son EQM est petite.

Un estimateur biaisé (c'est-à-dire de biais non nul) peut être plus précis qu'un estimateur sans biais.

Applications

Dans le cas général 1), on a : $\mathbb{E}(\overline{X}_n) = \mu$ et $\text{var}(\overline{X}_n) = \frac{\sigma^2}{n}$, donc $\overline{X}_n$ est un estimateur sans biais et convergent de μ.

On a $\mathbb{E}(S_n^2) = \sigma^2$ donc S_n^2 est un estimateur sans biais de σ^2 (mais S_n n'est pas un estimateur sans biais de σ).

Dans le cas Bernoulli 4), on a : $\mathbb{E}(F_n) = p$ et $\text{var}(F_n) = \frac{p(1-p)}{n}$, donc F_n est un estimateur sans biais et convergent de p.

3.4.3 Méthodes d'estimation

Méthode empirique

On sélectionne un estimateur naturel et on étudie ses propriétés. On le modifie éventuellement pour que le nouvel estimateur ait de meilleures propriétés (par exemple, la variance corrigée).

Méthode des moments

Si on a M paramètres à estimer $(\theta_m)_{m=1,..,M}$, on résout un système de M équations à M inconnues (les M paramètres) en égalant pour $m = 1, ..., M$, le moment d'ordre m de la population μ_m et le moment d'ordre m de l'échantillon $\frac{1}{n}\sum_{i=1}^{n}(X_i)^m$.

Exemple : l'estimateur du moment de σ^2 est alors la variance non corrigée de l'échantillon.

Méthode du maximum de vraisemblance

Soit $f(x, \theta)$ la loi de probabilité de X si elle est discrète (ou la densité de X si elle est continue).

La loi de probabilité conjointe (ou la densité conjointe) de $(X_1, ..., X_n)$ est alors :

$$L(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

C'est une fonction des observations $x_1, ..., x_n$ et du paramètre θ, appelée fonction de vraisemblance (likelihood en anglais, d'où la notation avec L).

La méthode du maximum de vraisemblance consiste à prendre pour θ la "valeur", fonction des $(x_i)_{i=1,...,n}$, qui maximise la vraisemblance, considérée comme fonction de θ. Pour utiliser cette méthode d'estimation, il faut donc connaître la loi de probabilité de X.

La fonction logarithme étant strictement croissante, maximiser la vraisemblance est équivalent à maximiser le logarithme de la vraisemblance.

Applications

Cas normal 3) : $X \sim N(\mu, \sigma)$, σ connu
$L(x_1, ..., x_n; \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2)$
$\ln L(x_1, ..., x_n; \mu) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$
$\frac{d\ln L}{d\mu} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i - \mu)$ $\quad$ $\frac{d\ln L}{d\mu} = 0 \Longleftrightarrow \mu = \overline{x}_n$
$\frac{d^2\ln L}{d\mu^2} = -\frac{n}{\sigma^2} < 0$, il s'agit bien d'un maximum.
L'estimateur par maximum de vraisemblance de μ est donc $\overline{X}_n$.

Cas normal 3) : $X \sim N(\mu, \sigma)$, μ connu
$\ln L(x_1, ..., x_n; \sigma) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$
$\frac{d\ln L}{d\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2$ $\quad$ $\frac{d\ln L}{d\sigma} = 0 \Longleftrightarrow \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2}$ $(= \sqrt{v_n})$
$\frac{d^2\ln L}{d\sigma^2} = \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$ $\quad$ $\frac{d^2\ln L}{d\sigma^2} < 0 \Longleftrightarrow \frac{\sigma^2}{3} < \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 = \sigma^2$
ce qui est vérifié, il s'agit bien d'un maximum.
L'estimateur par maximum de vraisemblance de σ est donc $\sqrt{v_n}$ et non s_n.)

Cas Bernoulli 4) : $X \sim \mathcal{B}(1, p)$, $P(X = x) = p$ si $x = 1$, $P(X = x) = 1 - p$ si $x = 0$.
$L(x_1, ..., x_n; p) = p^k(1 - p)^{n-k}$ où $k = \sum_{i=1}^{n} x_i$ (nombre de succès sur n épreuves).

$\ln L(x_1, ..., x_n; p) = k \ln p + (n-k)\ln(1-p)$
$\frac{d \ln L}{dp} = \frac{k}{p} - \frac{n-k}{1-p}$ $\quad$ $\frac{d \ln L}{dp} = 0 \Longleftrightarrow p = \frac{k}{n}$ ($= f_n$ fréquence de succès sur n épreuves).
$\frac{d^2 \ln L}{dp^2} = -\frac{k}{p^2} - \frac{n-k}{(1-p)^2} < 0$, il s'agit bien d'un maximum.

L'estimateur par maximum de vraisemblance de p est donc f_n.

4 Tests

4.1 Test d'égalité d'une moyenne à une valeur donnée

On se place dans le cas général : soit $X \sim (\mu, \sigma)$ et $(X_i)_i = 1, ..., n$ un échantillon de v.a.r. i.i.d. comme X (avec $n \geq 30$). On observe sur un échantillon de taille n la valeur $\overline{x}_n$ de $\overline{X}_n$ et on veut répondre à la question : peut-on accepter l'hypothèse que le paramètre inconnu μ soit égal à une valeur donnée μ_0.
Si l'hypothèse $\mu = \mu_0$ est vraie, alors $(\frac{\overline{X}_n - \mu_0}{s_n/\sqrt{n}}) \overset{approx}{\sim} N(0,1)$. Si l'on s'accorde un risque de se tromper de 5%, on en déduit :

$$P(\mu_0 - 1.96\frac{s_n}{\sqrt{n}} \leq \overline{X}_n \leq \mu_0 + 1.96\frac{s_n}{\sqrt{n}}) = 0.95$$

d'où la régle de décision :
- si $\overline{x}_n \in [\mu_0 - 1.96\frac{s_n}{\sqrt{n}} \ ; \ \mu_0 + 1.96\frac{s_n}{\sqrt{n}}]$, on "accepte" l'hypothèse (plus exactement, on ne la rejette pas), les données sont compatibles avec l'hypothèse ... ce qui ne veut pas dire qu'elle soit vraie ; on dit aussi que la différence entre la moyenne observée $\overline{x}_n$ et la moyenne théorique μ_0 n'est pas *significative au seuil de 5%.*

- sinon, on rejette l'hypothèse car on considère que la probabilité d'observer $\overline{x}_n$ sous cette hypothèse est trop faible (faible mais pas nulle, en fait, on rejette l'hypthèse avec une probabilité de se tromper de 5%) ; la différence entre $\overline{x}_n$ et μ_0 est dite *significative au seuil de 5%.*

Remarque

On a l'équivalence :

$$\overline{x}_n \in [\mu_0 - 1.96\frac{s_n}{\sqrt{n}} \ ; \ \mu_0 + 1.96\frac{s_n}{\sqrt{n}}] \quad \Longleftrightarrow \quad \mu_0 \in [\overline{x}_n - 1.96\frac{s_n}{\sqrt{n}} \ ; \ \overline{x}_n + 1.96\frac{s_n}{\sqrt{n}}].$$

Autrement dit, on rejette l'hypothèse au seuil de 5% lorsque la valeur de référence n'appartient pas à l'intervalle de confiance au seuil de 5% de la moyenne μ.

4.2 Test d'égalité d'une proportion à une valeur donnée

On se place dans le cas Bernoulli 4). Soit X une v.a.r. de Bernoulli de paramètre p et $(X_i)_i = 1, ..., n$ un échantillon de v.a.r. i.i.d. comme X (avec $n \geq 30$). On observe sur un échantillon de taille n la valeur f_n de la fréquence d'échantillonnage F_n et on veut répondre à la question : peut-on accepter l'hypothèse que le paramètre inconnu p soit égal à une valeur donnée p_0 ?
Si l'hypothèse $p = p_0$ est vraie, alors $(\frac{F_n - p_0}{\sqrt{p_0(1-p_0)/n}}) \sim N(0,1)$. Si l'on s'accorde un risque

de se tromper de 5%, on en déduit :

$$P(p_0 - 1.96\sqrt{\frac{p_0(1-p_0)}{n}} \leq F_n \leq p_0 + 1.96\sqrt{\frac{p_0(1-p_0)}{n}}) = 0.95$$

d'où la régle de décision :

- si $f_n \in [p_0 - 1.96\sqrt{\frac{p_0(1-p_0)}{n}}\ ;\ p_0 + 1.96\sqrt{\frac{p_0(1-p_0)}{n}}]$, on "accepte" l'hypothèse ; la différence entre f_n et p_0 n'est pas significative au seuil de 5%,

- sinon, on rejette l'hypothèse car on considère que la probabilité d'observer f_n sous cette hypothèse est trop faible ; la différence entre f_n et p_0 est significative au seuil de 5%.

Remarques

Contrairement au test sur la moyenne, pour une proportion, la variance dépend de la proportion ; aussi, la décision du test statistique ne se déduit pas de l'intervalle de confiance.

Les logiciels de statistique font évoluer la pratique des tests en fournissant pour les données observées une p-valeur (p-value en anglais) que l'on compare au seuil fixé à l'avance. Par exemple, dans le cas du test d'égalité d'une proportion à une valeur donnée p_0, si l'on observe une fréquence f_n sur un échantillon de taille n et donc un écart en valeur absolu $\delta = \mid f_n - p_0 \mid$, la p-valeur est $P(\mid F_n - p_0 \mid > \delta)$. Il s'agit de la probabilité, sous H_0, d'observer une "telle" différence. Si cette probabilité est plus petite que le seuil α fixé à l'avance, on rejette l'hypothèse $p = p_0$, sinon, on "accepte".

Les tests précédents sont appelés *tests paramétriques* car ils portent sur un paramètre de la distribution de probabilité de X. Les tests suivants, test du Khi2, sont des *tests non paramétriques*, construits sur les distributions de fréquence et de probabilité de ou des variables. Des tests non paramétriques, construits sur les rangs des valeurs des variables, permettent de s'affranchir d'hypothèse sur les distributions de probabilité et de travailler avec de petits échantillons. Au lieu de considérer les distributions asymptotiques des statistiques de test, on peut bien souvent utiliser les distributions de probabilité exactes.

4.3 Test du Khi2 d'adéquation d'une distribution de fréquence à une distribution de probabilité

Les § 4.3 et 4.4 reprennent le document d'accompagnement des programmes de terminale S et ES, page 145.

Le dé est-il équilibré ?

Supposons que sur n lancers d'un dé, on observe la distribution de fréquences $(f_n^1, ..., f_n^6)$.

Si le dé est équilibré, ce vecteur est asymptotiquement égal au vecteur $(\frac{1}{6}, ..., \frac{1}{6})$ et, si on calcule :

$$a_n^2 = n\sum_{h=1}^{6}\frac{(f_n^h - \frac{1}{6})^2}{\frac{1}{6}} = 6n\sum_{h=1}^{6}(f_n^h - \frac{1}{6})^2 = n(6\sum_{h=1}^{6}(f_n^h)^2 - 1),$$

alors a_n^2 est l'observation d'une v.a.r. A_n^2 qui suit asymptotiquement un $Khi2$ à 5 degrés de liberté.

On lit sur la table des lois de $Khi2$: $P(\chi^2_5 \leq 9.24) = 0.90$.

Règle de décision :

- si $a_n^2 > 9.24$, on rejette l'hypothèse que le dé soit équilibré (avec une probabilité d'erreur de 10%),

- si $a_n^2 \leq 9.24$, on ne rejette pas l'hypothèse que le dé soit équilibré (les données ne nous permettent pas de rejeter cette hypothèse).

Dans le document d'accompagnement des programmes de terminale S et ES, on pose $d^2 = \sum_{h=1}^{6}(f_n^h - \frac{1}{6})^2$ et on réalise tout d'abord 2000 simulations avec $n = 500$, on observe que 90% des d^2 observés sont compris entre 0 et 0.0030, c'est-à-dire que 90% des a_n^2 $(= 6 \times 500 \times d^2)$ sont compris entre 0 et 9.

On retrouve par simulation le résultat théorique $P(\chi^2_5 \leq 9.24) = 0.90$.

On réalise ensuite 5000 simulations avec $n = 1000$, on observe que 90% des d^2 observés sont compris entre 0 et 0.0015, c'est-à-dire que 90% des a_n^2 $(= 6 \times 1000 \times d^2)$ sont compris entre 0 et 9.

On retrouve encore par simulation le résultat théorique concernant un $Khi2$ à 5 degrés de liberté.

4.4 Test du Khi2 d'indépendance de deux variables aléatoires catégorielles

Il s'agit du deuxième problème du document d'accompagnement.

Enquête marketing

Une enquête marketing portant sur le choix entre deux abonnements A et B lors de l'achat d'un téléphone portable et le statut de l'acheteur (salarié ou non) a conduit au recueil des données sur 9321 nouveaux acheteurs enregistrés sur un fichier client (l'étude portait sur 10000 acheteurs mais pour 679 d'entre eux, on ne disposait pas du statut).

On a obtenu les distributions d'effectifs et de fréquences suivantes :

	A	B	Ens
S	4956	1835	6791
NS	1862	668	2530
Ens	6818	2503	9321

	A	B	Ens
S	0.532	0.197	0.729
NS	0.200	0.071	0.271
Ens	0.732	0.268	1

Si les variables "choix de l'abonnement" et "statut" étaient indépendantes, on observerait, en supposant les marges fixées, la distribution de fréquences théoriques suivante (tableau de proportionnalité) :

	A	B	Ens
S	0.534	0.195	0.729
NS	0.198	0.073	0.271
Ens	0.732	0.268	1

La distance du χ^2 entre la distribution de fréquences observée et la distribution de fréquences théoriques d'indépendance est alors :

$$\chi^2 = 9321(\frac{0.532^2}{0.534} + \frac{0.197^2}{0.195} + \frac{0.200^2}{0.198} + \frac{0.071^2}{0.073} - 1) = 0.96.$$

Sous l'hypothèse d'indépendance, il s'agit de l'observation d'un χ^2 à 1 degré de liberté ($(p-1)(q-1)$ si p et q sont les nombres de modalités des deux variables).

On lit dans les tables de χ^2 : $P(\chi^2_{1ddl} < 2.71) = 0.90$.

On "accepte" l'hypothèse d'indépendance des variables (la valeur observée est compatible avec l'hypothèse).

On a construit la distribution de fréquences théoriques sous hypothèse d'indépendance en supposant les marges fixées. En fait, on utilise ainsi une estimation de la répartition de 9321 individus selon les modalités de chacune des deux variables.

4.5 Compléments théoriques sur les tests : procédure générale d'un test

La procédure générale d'un test est la suivante :

1) On définit le modèle statistique en précisant ce qui est connu et ce qui ne l'est pas.
Par exemple : $(X_i)_{i=1,\dots,n}$ i.i.d. comme $X \sim N(\mu, \sigma)$ avec $\sigma = 1$.

2) On définit l'hypothèse à tester H_0, appelée *hypothèse nulle*, et l'*hypothèse alternative*, notée H_1.
Par exemple : $H_0 : \mu = 24$, $H_1 : \mu \neq 24$

Dans cet exemple, H_0 est une *hypothèse simple* (car de la forme "le paramètre appartient à un singleton") alors que l'hypothèse alternative est une *hypothèse composite* (car de la forme "le paramètre appartient à un ensemble d'au moins deux éléments"). Dans toute la suite, nous ne traiterons que le cas où l'hypothèse nulle est simple (donc de la forme $\mu = \mu_0$).

Le test de cet exemple est *bilatéral* car l'hypothèse alternative envisage des valeurs du paramètre de part et d'autre de μ_0. Le test $H_0 : \mu = 24$, $H_1 : \mu < 24$ (resp. $H_0 : \mu = 24$, $H_1 : \mu > 24$) est un test *unilatéral*.

Remarque : le test $H_0 : \mu \geq 24$, $H_1 : \mu < 24$ (resp. $H_0 : \mu \leq 24$, $H_1 : \mu > 24$) se ramène au test $H_0 : \mu = 24$, $H_1 : \mu < 24$ (resp. $H_0 : \mu = 24$, $H_1 : \mu > 24$).

3) On utilise un estimateur du paramètre ou d'une fonction du paramètre dont on connaît la loi de probabilité sous H_0 (c'est-à-dire, si H_0 est vraie).
Par exemple : $\overline{X}_n \sim N(24, 1)$ sous H_0

4) On se fixe un seuil α, appelé *risque de 1ère espèce*, correspondant à une probabilité d'erreur que l'on s'accorde, par exemple : $\alpha = 5\%$.

5) On définit la *région de rejet* (appelée aussi *région critique*) et la *région d'acceptation* correspondant à ce seuil. Les valeurs appartenant à la frontière entre les deux régions sont appelées *valeurs critiques.*
Par exemple, pour le test bilatéral : $H_0 : \mu = 24$, $H_1 : \mu \neq 24$, on a :
$P(24 - 1.96 \leq \overline{X}_n \leq 24 + 1.96) = 0.95$.
Les valeurs critiques sont 22.04 et 25.96, la région de rejet est $]-\infty, 22.04[\cup]25.96, \infty[$ et la région d'acceptation est $[22.04, 25.96]$.
Pour le test unilatéral : $H_0 : \mu = 24$, $H_1 : \mu < 24$, on a : $P(24 - 1.645 \leq \overline{X}_n) = 0.95$.

La valeur critique est 22.355, la région de rejet $]-\infty, 22.355[$, la région d'acceptation $[22.355, \infty[$.

6) On énonce la *règle de décision* :

Soit $\overline{x}_n$ une observation de $\overline{X}_n$: on rejette H_0 (et donc on accepte H_1) si $\overline{x}_n$ appartient à la région critique ; on accepte H_0 si $\overline{x}_n$ appartient à la région d'acceptation. En fait, plutôt que de dire "on accepte H_0" on devrait dire "on ne rejette pas H_0". En effet, on conclut seulement que les observations faites ne sont pas en contradiction avec l'hypothèse H_0.

7) *Erreurs de 1ère et de 2ème espèce, puissance du test* et *courbe d'efficacité.*

Le tableau suivant fait apparaître deux erreurs possibles. On a, en ligne, "l'état du phénomène observé" : H_0 vraie ou bien H_1 vraie, et, en colonne, la décision prise "on accepte H_0" ou bien "on rejette H_0". L'erreur de première espèce correspond au cas où on rejette H_0 alors qu'elle est vraie. L'erreur de deuxième espèce correspond au cas où on "accepte" H_0 alors qu'elle est fausse.

Décision \ *État*	H_0 *vraie*	H_1 *vraie*
on accepte H_0	/	*erreur* 2° *esp.*
on rejette H_0	*erreur* 1° *esp.*	/

- le seuil α que l'on s'est fixé est la probabilité sous H_0 de rejeter H_0. Cette probabilité est appelée *risque de 1ère espèce* ou risque α.
- la probabilité sous H_1 d'accepter H_0 est appelée *risque de 2ème espèce* ou risque β.

La *puissance du test* est par définition égale à $1-\beta$.

On pourra calculer le risque de deuxième espèce et la puissance du test dans le cas où l'hypothèse alternative est simple. Par exemple : $H_0 : \mu = 24$, $H_1 : \mu = 22$.

La détermination de la région de rejet se fait comme pour le test unilatéral : $H_0 : \mu = 24$, $H_1 : \mu < 24$.

On cherche à présent, sous H_1 (c'est-à-dire $\overline{X}_n \sim N(22, 1)$) la probabilité que $\overline{X}_n$ appartienne à la région d'acceptation. On a : $\beta = P(\overline{X}_n \geq 22.355) = P(U_n \geq 0.355) = 1 - P(U_n \leq 0.355) = 0.361$.

Dans le cas où l'hypothèse alternative est composite, par exemple $H_0 : \mu = 24$, $H_1 : \mu \neq 24$, *la courbe d'efficacité* est le graphe de la puissance du test fonction des valeurs du paramètre $\beta(\mu)$.

Diminuer α entraîne une augmentation de β et donc une diminution de la puissance du test. Il s'agit donc de chercher un compromis afin que les deux risques soient suffisamment faibles.