Formules pour l'examen 1

Formules et définitions générales :

Théorème de Bayes :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^C)P(A^C)}$$

où A^C est le complémentaire de l'événement A, donc $P(A^C) = 1 - P(A)$.

Quantile : Soit $W \sim \mathcal{L}$, où $\mathcal{L}$ est une loi quelconque. Le quantile w_α de cette loi est définit comme étant la valeur vérifiant l'équation $P(W > w_\alpha) = \alpha$.

Définitions relatives aux tests d'hypothèse :

H_0 **:** hypothèse nulle ;

H_1 **:** hypothèse alternative ;

Seuil ou niveau de signification, noté α : probabilité de commettre une erreur de type I, soit $P(\text{rejeter } H_0 \mid H_0 \text{ est vraie})$.

Seuil observé (en anglais p-value) : probabilité, sous H_0, d'obtenir un résultat égal ou plus extrême que celui observé ;

mid p-value : la moitié de la probabilité d'un résultat aussi probable que celui observé, plus la probabilité d'un résultat plus extrême (moins probable), sous H_0.

Définitions relatives à la vraisemblance d'un vect. de param. $\boldsymbol{\theta}$:

Vraisemblance : Soit $X_1, \ldots, X_n$ un échantillon aléatoire de taille n, de fonction de masse ou de densité conjointe $f(\boldsymbol{x}|\boldsymbol{\theta})$. Conditionnellement à ce que $\boldsymbol{X} = (X_1, \ldots, X_n)$ soit observé et prenne la valeur $\boldsymbol{x} = (x_1, \ldots, x_n)$, la fonction de vraisemblance est définie par :

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}).$$

Si les variables aléatoires X_1 à X_n sont indépendantes, on a que

$$L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{l=1}^{n} f(x_l|\boldsymbol{\theta}).$$

Fonction score :
vecteur des dérivées partielles de la log-vraisemblance par rapport aux paramètres :

$$S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}|\boldsymbol{x}).$$

Matrice d'information espérée :
(aussi appelée matrice d'information de Fisher ou simplement matrice d'information) :

$$I(\boldsymbol{\theta}) = E\left(S(\boldsymbol{\theta})^2\right) = E\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{X}|\boldsymbol{\theta})\right)^2\right)$$

Si X suit une distribution de le famille exponentielle, cette matrice se simplifie à :

$$I(\boldsymbol{\theta}) = -E\left(\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ln f(\boldsymbol{X}|\boldsymbol{\theta})\right).$$

Cette matrice est en fait la matrice de variance-covariance de la statistique score.

Estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$, noté $\hat{\boldsymbol{\theta}}$:
point en lequel la vraisemblance $L(\boldsymbol{\theta}|\boldsymbol{x})$ est maximisée.

Tests asymptotiques usuels :

Statistique d'un test de Wald sur un paramètre θ :
$H_0 : \theta = \theta_0$, H_1 peut être bilatéral ($\theta \neq \theta_0$) ou unilatéral ($\theta > \theta_0$ ou $\theta < \theta_0$)

$$\frac{\theta^* - \theta_0}{se(\theta^*)} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1)$$

où θ^* est un estimateur ponctuel de θ et $se(\theta^*)$ est un estimateur de l'erreur type de θ^*.

Statistique d'un test score sur un paramètre θ :
$H_0 : \theta = \theta_0$, H_1 peut être bilatéral ($\theta \neq \theta_0$) ou unilatéral ($\theta > \theta_0$ ou $\theta < \theta_0$)

$$\frac{S(\theta_0)}{\sqrt{I(\theta_0)}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad N(0,1)$$

où $S(\theta_0)$ est la fonction score calculée au point $\theta = \theta_0$ et $I(\theta_0)$ est la matrice d'information espérée (ici de dimension 1×1) calculée au point $\theta = \theta_0$.

Statistique d'un test de rapport de vraisemblance sur un vect. de param. $\boldsymbol{\theta}$:
$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, H_1 peut seulement être bilatérale ($\boldsymbol{\theta} \in \boldsymbol{\Theta}/\boldsymbol{\Theta}_0$) où $\boldsymbol{\Theta}_0$ représente un sous-ensemble de l'espace $\boldsymbol{\Theta}$ des valeurs possible de $\boldsymbol{\theta}$

$$-2 \ln \Lambda(\boldsymbol{X}) \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_d.$$

où

$$\Lambda(\boldsymbol{x}) = \frac{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} L(\boldsymbol{\theta}|\boldsymbol{x})}{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta}|\boldsymbol{x})} = \frac{L(\hat{\boldsymbol{\theta}}_0|\boldsymbol{x})}{L(\hat{\boldsymbol{\theta}}|\boldsymbol{x})}$$

avec $\hat{\boldsymbol{\theta}}$ = l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ et
$\hat{\boldsymbol{\theta}}_0$ = l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ restreint sous l'espace $\boldsymbol{\Theta}_0$.
Le nombre de degrés de liberté d est la différence entre le nombre de paramètres libres sous l'espace $\boldsymbol{\Theta}$ (en d'autres mots la dimension de $\boldsymbol{\Theta}$) et le nombre de paramètres libres sous $\boldsymbol{\Theta}_0$ (en d'autres mots la dimension de $\boldsymbol{\Theta}_0$).

Définitions relatives aux intervalles de confiance :

Niveau de confiance, noté $1-\alpha$ **:** probabilité que le paramètre soit inclut dans l'intervalle de confiance, i.e. $P(L \leq \theta \leq U) = 1-\alpha$.

Intervalle de confiance de Wald d'un paramètre θ **:** $[\theta^* - z_{\alpha/2}se(\theta^*), \theta^* + z_{\alpha/2}se(\theta^*)]$ où θ^* est un estimateur ponctuel de θ et $se(\theta^*)$ un estimateur de l'erreur type de θ^*

1 Formules relatives aux tableaux de fréquences à une variable

Résumé des informations relatives aux trois distributions étudiées :

	Binomiale	**Poisson**	**Multinomiale**
Échantillon sur l'échelle catégorique	Y_1 à $Y_n \in \{\text{succès}, \text{échec}\}$	-	Y_1 à $Y_n \in \{res_1, \ldots, res_k\}$
Échantillon sur l'échelle numérique	$X_l = \begin{cases} 1 & \text{si } Y_l = \text{succès} \\ 0 & \text{si } Y_l = \text{échec} \end{cases}$ pour $l = 1, \ldots, n$	X_1 à $X_n \in \{0, 1, 2, \ldots\}$	$X_{il} = \begin{cases} 1 & \text{si } Y_l = res_i \\ 0 & \text{sinon} \end{cases}$ pour $i = 1, \ldots, k$ et $l = 1, \ldots, n$
Définition de la ou des variables aléatoires	une seule v.a. $X = \sum_l X_l$: nombre de succès parmi les n individus de l'échantillon	n v.a. indépendantes X_l : nombre de réalisations d'un événement dans un intervalle de temps et/ou d'espace pour chaque individu de l'échantillon	un vecteur de k v.a. $\boldsymbol{X} = (X_1, \ldots, X_k)$ avec $X_i = \sum_l X_{il}$: nombre d'individus de l'échantillon pour lesquels $Y = res_i$, pour $i = 1, \ldots, k$
Tableau des fréquences observées	$\begin{array}{c\|cc\|c} Y & \text{succès} & \text{échec} & \text{total} \\ \hline \text{fréq.} & x & n-x & n \end{array}$	$\begin{array}{c\|cccc\|c} X & 0 & 1 & \cdots & & \text{total} \\ \hline \text{fréq.} & n_0 & n_1 & \cdots & & n \end{array}$	$\begin{array}{c\|ccc\|c} Y & res_1 & \cdots & res_k & \text{total} \\ \hline \text{fréq.} & x_1 & \cdots & x_k & n \end{array}$
Notation	$X \sim Bin(n, \pi)$	X_l *iid* $Poisson(\lambda)$ pour $l = 1, \ldots, n$	$\boldsymbol{X} = (X_1, \ldots, X_k) \sim Multinomiale(n, \pi_1, \ldots, \pi_k)$
valeurs possibles	$\{0, \ldots, n\}$	$\{0, 1, 2, \ldots\}$	$\{0, \ldots, n\}$ $\forall X_i$ sous la contrainte que $X_1 + \ldots + X_k = n$

	Binomiale	**Poisson**	**Multinomiale**
Fonction de masse	$P(X=x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}$	$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$	$P(\boldsymbol{X}=(x_1,\ldots,x_k)) = \frac{n!}{n_1!\ldots n_k!}\pi_1^{x_1}\ldots\pi_k^{x_k}$
Espérance	$E(X) = n\pi$	$E(X) = \lambda$	$E(X_i) = n\pi_i$ pour $i = 1,\ldots,k$
Variance	$Var(X) = n\pi(1-\pi)$	$Var(X) = \lambda$	$Var(X_i) = n\pi_i(1-\pi_i)$ pour $i = 1,\ldots,k$ et $Cov(X_i, X_{i'}) = -n\pi_i\pi_{i'}$ pour $i \neq i'$
Paramètre d'intérêt	$\pi = P(Y = \text{succès})$	$\lambda = E(X)$	$\boldsymbol{\pi} = (\pi_1,\ldots\pi_k)$ où $\pi_i = P(Y = res_i)$, sous la contrainte $\sum_i \pi_i = 1$
Estimateur max. vrais. du paramètre	$\hat{\pi} = \frac{X}{n}$	$\hat{\lambda} = \frac{\sum_{l=1}^n X_l}{n}$	$\hat{\pi}_i = \frac{X_i}{n}$ pour $i = 1,\ldots,k$
Hypothèses d'un test sur le paramètre	$H_0 : \pi = \pi_0$ $H_1 : \pi \neq$ ou $>$ ou $< \pi_0$	$H_0 : \lambda = \lambda_0$ $H_1 : \lambda \neq$ ou $>$ ou $< \lambda_0$	$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi_0} = (\pi_{0,1},\ldots,\pi_{0,k})$ $H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi_0}$
Statistique du test de Wald	$Z_w = \frac{\hat{\pi}-\pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	$Z_w = \frac{\hat{\lambda}-\lambda_0}{\sqrt{\hat{\lambda}/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	-
Statistique du test score	$Z_s = \frac{\hat{\pi}-\pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	$Z_w = \frac{\hat{\lambda}-\lambda_0}{\sqrt{\lambda_0/n}} \xrightarrow[H_0]{\text{asymp.}} \mathcal{N}(0,1)$	$X^2 = \sum_{i=1}^k \frac{(X_i - n\pi_{0,i})^2}{n\pi_{0,i}} \xrightarrow[H_0]{\text{asymp.}} \chi^2_{k-1}$
Statistique test rapport vraisemblance (test bilatéral seulement)	$W_{rv} = -2\left(x\ln\left(\frac{\pi_0}{\hat{\pi}}\right) + (n-x)\ln\left(\frac{1-\pi_0}{1-\hat{\pi}}\right)\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_1$	$W_{rv} = -2n\left(\hat{\lambda}\ln\left(\frac{\lambda_0}{\hat{\lambda}}\right) + (\hat{\lambda}-\lambda_0)\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_1$	$G^2 = -2\sum_{i=1}^k X_i \ln\left(\frac{\pi_{0,i}}{\hat{\pi}_i}\right) \xrightarrow[H_0]{\text{asymp.}} \chi^2_{k-1}$
Statistique du test exact	$X \underset{H_0}{\longrightarrow} Bin(n, \pi_0)$	-	-
IC Wald de niveau $1-\alpha$	$\hat{\pi} \pm z_{\alpha/2}\sqrt{\hat{\pi}(1-\hat{\pi})/n}$	$\hat{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}$	correction de Bonferroni : IC individuels de niveau de confiance $1-\alpha/k$ pour les k π_i comme pour le paramètre de la binomiale
IC score de niveau $1-\alpha$	$[L_\pi, U_\pi]$ tel que définit sous le tableau	$[L_\lambda, U_\lambda]$ tel que définit sous le tableau	

$$L_\pi = \frac{n}{n+z^2_{\alpha/2}}\left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}\right) \text{ et } U_\pi = \frac{n}{n+z^2_{\alpha/2}}\left(\hat{\pi} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n} + \frac{z^2_{\alpha/2}}{4n^2}}\right),$$

$$L_\lambda = \hat{\lambda} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n} + \frac{z^2_{\alpha/2}}{4n^2}} \text{ et } U_\lambda = \hat{\lambda} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{\lambda}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}.$$

Remarque : Pour trouver les valeurs observées des estimteurs et des statistiques de test, il suffit de remplacer les variables aléatoires X, X_l ou X_i par leurs valeurs observées x, x_l ou x_i, respectivement.

Formules de statistiques descriptives numériques selon le format des données :

Statistique	Format individus	Format fréquences
moyenne ($\bar{x}$)	$\frac{\sum_{l=1}^{n} x_l}{n}$	$\frac{\sum_{i=1}^{k} v_i n_i}{n}$
variance	$\frac{\sum_{l=1}^{n}(x_l-\bar{x})^2}{n-1} = \frac{\sum_{l=1}^{n} x_l^2 - n\bar{x}^2}{n-1}$	$\frac{\sum_{i=1}^{k} v_i^2 n_i - n\bar{x}^2}{n-1}$

Formes générales pour les statistiques du khi-deux de Pearson (X^2) et du rapport de vraisemblance (G^2) :

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad \text{et} \qquad G^2 = 2\sum_{i=1}^{k} O_i \ln\left(\frac{O_i}{E_i}\right)$$

où les O_j sont des fréquences observées et les E_j sont des fréquences espérées. Sous l'hypothèse nulle que les fréquences espérées soient vraies, les statistiques X^2 et G^2 suivent asymptotiquement une loi du khi-deux à d degrés de liberté. Ces degrés de liberté sont la différence entre le nombre de paramètres libres dans l'espace de toutes les valeurs possibles des fréquences O_1 à O_k et le nombre de paramètres libres sous l'hypothèse nulle.

Remarque : La validité de loi asymptotique des statistiques X^2 et G^2 peut être mise en doute lorsque plus de 20% des fréquences espérées sont inférieures à 5.

Test d'adéquation de données à une loi avec les statistiques X^2 et G^2 :
Étape préalable : déterminer arbitrairement k classes qui couvrent tout le support des valeurs possible de la variable à tester. Pour $i = 1, \ldots, k$, O_i est le nombre d'observations de l'échantillon tombant dans la classe k.
Selon le type de l'hypothèse nulle, les statistiques de test sont les suivantes :

type 1 $\rightarrow$ H_0 : La loi $\mathcal{L}(\theta_0)$ s'ajuste bien aux données
X^2 ou $G^2 \xrightarrow[H_0]{\text{asymp.}} \chi^2_{k-1}$
avec $E_i = nP(Y \in \text{ classe } i | Y \sim \mathcal{L}(\boldsymbol{\theta}_0))$

type 2 $\rightarrow$ H_0 : La famille de loi $\mathcal{L}$ s'ajuste bien aux données
X^2 ou $G^2 \xrightarrow[H_0]{\text{asymp.}} \chi^2_{k-1-p}$
avec $E_i = nP(Y \in \text{ classe } i | Y \sim \mathcal{L}(\hat{\boldsymbol{\theta}}))$

où p est le nombre de paramètres de la loi $\mathcal{L}$ que l'on estime à partir des données et $\hat{\boldsymbol{\theta}}$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$.

2 Tableaux de fréquences à 2 variables

2.1 Définitions

Variable en lignes : X, indice : $i = 1, \ldots, I$, modalités : v_i (parfois variable explicative) ;
Variable en colonnes : Y, indice : $j = 1, \ldots, J$, modalités : w_j (parfois variable réponse).

Fréquences théoriques :

		Y			
		v_1	$\cdots$	v_J	total
X	u_1				
	$\vdots$		μ_{ij}		$\mu_{i\bullet}$
	u_I				
	total		$\mu_{\bullet j}$		n

Fréquences observées :

		Y			
		v_1	$\cdots$	v_J	total
X	u_1				
	$\vdots$		n_{ij}		$n_{i\bullet}$
	u_I				
	total		$n_{\bullet j}$		n

Tous les estimateurs et les statistiques seront ici énoncés en utilisant les fréquences théoriques. Cette notation met en évidence le fait qu'il s'agit de variables aléatoires (fonctions d'un échantillon aléatoire). Pour un échantillon donné, on calcule les valeurs observées de ces quantités en remplaçant les μ_{ij}, $\mu_{i\bullet}$ et $\mu_{\bullet j}$ par n_{ij}, $n_{i\bullet}$ et $n_{\bullet j}$, respectivement.

Types de fréquences :

Fréquences croisées : μ_{ij} ;

Fréquences marginales : $\mu_{i\bullet}$ et $\mu_{\bullet j}$;

Fréquences conditionnelles : Une ligne ou une colonne de fréquences croisées ;

Fréquences relatives croisées : μ_{ij}/n ;

Fréquences relatives marginales : $\mu_{i\bullet}/n$ et $\mu_{\bullet j}/n$;

Fréquences relatives conditionnelles :
fréquences de X conditionnelles à Y : $\mu_{ij}/n_{\bullet j}$
fréquences de Y conditionnelles à X : $\mu_{ij}/n_{i\bullet}$.

Types d'échantillonnage :

multinomial vs Poisson :
- multinomial : la taille d'échantillon n est fixe, interprétation statistique :

$$(\mu_{ij}, i = 1, \ldots, I; j = 1, \ldots, J) \sim$$
$$Multinomiale(n, \pi_{ij}, i = 1, \ldots, I; j = 1, \ldots, J).$$

- Poisson : la taille d'échantillon n n'est pas fixe, interprétation statistique :

$$\mu_{ij} \sim Poisson(\lambda_{ij}) \qquad \text{indépendantes pour} \quad i = 1, \ldots, I \quad \text{et} \quad j = 1, \ldots, J$$

simple vs multiple :
- simple : un seul échantillon (comme ci-dessus)

– multiple : On forme des sous-populations (strates), à partir des modalités de la variable X ou de la variable Y, et on tire un échantillon dans chacune des sous-populations. Ces échantillons sont indépendants. Si l'échantillonnage est multinomial, on a l'interprétation statistique suivante :
Stratification par rapport à la X : on a I sous-populations indépendantes telles que :

$$(\mu_{i1}, \ldots, \mu_{iJ}) \sim Multinomiale(n_i, \pi_{1|i}, \ldots, \pi_{J|i}) \quad \text{pour } i = 1, \ldots, I$$

où les n_i sont les $n_{i\bullet}$ vus auparavant, mais considérés fixes.
Stratification par rapport à Y : on a J sous-populations indépendantes telles que :

$$(\mu_{1j}, \ldots, \mu_{Ij}) \sim Multinomiale(n_j, \pi_{1|j}, \ldots, \pi_{I|j}) \quad \text{pour } j = 1, \ldots, J$$

où les n_j sont les $n_{\bullet j}$ vus auparavant, mais considérés fixes.

Probabilités d'intérêt et leurs estimateurs :

Type de probabilité	Probabilité	Définition	Estimateur potentiel	Bon estimateur si éch. multiple
conjointe	π_{ij}	$P(X = v_i, Y = w_j)$	μ_{ij}/n	jamais
marginale	$\pi_{i\bullet}$	$P(X = v_i)$	$\mu_{i\bullet}/n$	si var. stratif. $= Y$
	$\pi_{\bullet j}$	$P(Y = w_j)$	$\mu_{\bullet j}/n$	si var. stratif. $= X$
conditionnelle	$\pi_{i\|j}$	$P(X = v_i \| Y = w_j)$	$\mu_{ij}/n_{\bullet j}$	si var. stratif. $= Y$
	$\pi_{j\|i}$	$P(Y = w_j \| X = v_i)$	$\mu_{ij}/n_{i\bullet}$	si var. stratif. $= X$

2.2 Test d'association entre deux variables nominales

Test d'indépendance :

$$\begin{aligned} H_0 \quad : \quad & X \text{ et } Y \text{ sont indépendants} \quad \text{ou} \\ & \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \quad \forall\ i,\ j \end{aligned}$$

Test d'homogénéité de sous-populations (stratification par rapport à X) :

$$\begin{aligned} H_0 \quad : \quad & \text{Dans les } I \text{ sous-populations déterminées par } X, \\ & Y \text{ suit la même distribution} \quad \text{ou} \\ & (\pi_{11}, \ldots, \pi_{1J}) = \cdots = (\pi_{I1}, \ldots, \pi_{IJ}) \quad \text{ou} \\ & \pi_{j|i} = \pi_{j|i'} \quad \forall\ i \neq i',\ j \quad \text{ou} \\ & \pi_{j|i} = \pi_{\bullet j} \quad \forall\ i,\ j \end{aligned}$$

Deux tests équivalents car : indépendance $\Leftrightarrow$ homogénéité des sous-populations.

Les hypothèse alternatives sont le complément des hypothèses nulles, ces tests sont toujours bilatéraux.

– **Statistique du khi-deux de Pearson** :

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\mu_{ij} - \mu_{ij}^{H_0})^2}{\mu_{ij}^{H_0}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\mu_{ij} - \mu_{i\bullet}\mu_{\bullet j}/n)^2}{\mu_{i\bullet}\mu_{\bullet j}/n} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_{(I-1)(J-1)}$$

– **Statistique du rapport de vraisemblance** :

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \mu_{ij} \ln \frac{\mu_{ij}}{\mu_{ij}^{H_0}} = 2\sum_{i=1}^{I}\sum_{j=1}^{J} \mu_{ij} \ln \frac{\mu_{ij}}{\mu_{i\bullet}\mu_{\bullet j}/n} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_{(I-1)(J-1)}$$

Cas particulier des tableaux 2 × 2 :

Test de comparaison de deux proportions :
Posons $\pi_1 = \pi_{1|i=1}$ et $\pi_2 = \pi_{1|i=2}$.

$$H_0 : \pi_1 = \pi_2 \qquad \text{versus} \qquad H_1 : \pi_1 \neq \text{ ou } > \text{ ou } < \pi_2$$

– **Statistique du test de Wald** :

$$Z_w = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\dfrac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1\bullet}} + \dfrac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2\bullet}}}} \quad \xrightarrow[H_0]{\text{asymp}} \quad \mathcal{N}(0,1).$$

– **Statistique du test score** :

$$Z_s = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left(\dfrac{1}{n_{1\bullet}} + \dfrac{1}{n_{2\bullet}}\right)}} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \mathcal{N}(0,1)$$

où $\hat{\pi}_1 = \mu_{11}/n_{1\bullet}$, $\hat{\pi}_2 = \mu_{21}/n_{2\bullet}$ et $\hat{\pi} = \frac{n_{1\bullet}\hat{\pi}_1 + n_{2\bullet}\hat{\pi}_2}{n_{1\bullet}+n_{2\bullet}} = \frac{\mu_{11}+\mu_{21}}{n}$.

Forme abrégée de la statistique du khi-deux de Pearson pour un tableau 2 × 2 :

$$X^2 = \frac{n\Delta^2}{\mu_{1\bullet}\mu_{2\bullet}\mu_{\bullet 1}\mu_{\bullet 2}} \qquad \text{où} \quad \Delta = \mu_{11}\mu_{22} - \mu_{12}\mu_{21}$$

Petits échantillons :

Correction pour la continuité de la statistique du khi-deux de Pearson pour un tableau 2 × 2 (correction de Yates) :

$$X^2_{corr} = \frac{n(|\Delta| - \frac{n}{2})^2}{\mu_{1\bullet}\mu_{2\bullet}\mu_{\bullet 1}\mu_{\bullet 2}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_1$$

Test exact de Fisher pour un tableau 2 × 2 :
Test d'indépendance dont la statistique de test est :

$$\mu_{11} \quad \underset{H_0}{\longrightarrow} \quad Hypergéométrique(a = n_{\bullet 1}, b = n_{1\bullet}, c = n)$$

Cette distribution est exacte, mais elle suppose que les marges sont fixes.

Fonctions de masse de la distribution $Hypergéométrique(a, b, c)$:

$$P(X = x) = \frac{\binom{b}{x}\binom{c-b}{a-x}}{\binom{c}{a}} \qquad \text{pour} \quad max(0, a + b - c) \leq x \leq min(a, b)$$

2.3 Décrire et mesurer l'association entre deux variables nominales

– **Probabilités conditionnelles** : telles que définies à la section 2.1.

– **Résidus** :
  – **Résidus bruts** :
$$RB_{ij} = \mu_{ij} - \mu_{ij}^{H_0} = \mu_{ij} - \mu_{i\bullet}\mu_{\bullet j}/n$$
  – **Résidus de Pearson** :
$$RP_{ij} = \frac{\mu_{ij} - \mu_{i\bullet}\mu_{\bullet j}/n}{\sqrt{\mu_{i\bullet}\mu_{\bullet j}/n}} \qquad \underset{H_0}{\xrightarrow{\text{asymp.}}} \qquad N(0, \sigma_{ij}^2)$$
  – **Résidus de Pearson ajustés ou standardisés** :
$$RAP_{ij} = \frac{\mu_{ij} - \mu_{i\bullet}\mu_{\bullet j}/n}{\sqrt{\frac{\mu_{i\bullet}\mu_{\bullet j}}{n}\left(1 - \frac{\mu_{i\bullet}}{n}\right)\left(1 - \frac{\mu_{\bullet j}}{n}\right)}} \qquad \underset{H_0}{\xrightarrow{\text{asymp.}}} \qquad N(0, 1)$$

– **Coefficient de Cramer** :

$$V = \sqrt{\frac{X^2/n}{\min(I-1, J-1)}} \qquad \text{pour un tableau } 2 \times 2 : V = \frac{\mu_{11}\mu_{22} - \mu_{12}\mu_{21}}{\sqrt{\mu_{1\bullet}\mu_{2\bullet}\mu_{\bullet 1}\mu_{\bullet 2}}}$$

– **Différence de proportions** pour un tableau 2 × 2 :
Définition théorique :
$$\pi_{1|i=1} - \pi_{1|i=2} = \pi_1 - \pi_2$$

Estimateur :
$$\hat{\pi}_1 - \hat{\pi}_2 = \mu_{11}/n_{1\bullet} - \mu_{21}/n_{2\bullet}$$

Intervalle de confiance de niveau $1-\alpha$:

$$\pi_1 - \pi_2 \ \epsilon \ \hat{\pi}_1 - \hat{\pi}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_{1\bullet}} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_{2\bullet}}}$$

– **Risque relatif** pour un tableau 2×2 :
Définition théorique :

$$RR = \frac{\pi_1}{\pi_2}$$

Estimateur :

$$\widehat{RR} = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{\mu_{11}/\mu_{1\bullet}}{\mu_{21}/\mu_{2\bullet}}$$

Intervalle de confiance de niveau $1-\alpha$:

$$\left[\widehat{RR}\ e^{-z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))},\ \widehat{RR}\ e^{z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RR}))}\right] \quad \text{avec } \hat{\sigma}(\ln(\widehat{RR})) = \sqrt{\frac{1}{\mu_{11}} - \frac{1}{n_{1\bullet}} + \frac{1}{\mu_{21}} - \frac{1}{n_{2\bullet}}}$$

– **Rapport de cotes** pour un tableau 2×2 :
Définition théorique :

$$RC = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$$

Estimateur :

$$\widehat{RC} = \frac{\hat{\pi}_{11}\hat{\pi}_{22}}{\hat{\pi}_{21}\hat{\pi}_{12}} = \frac{\mu_{11}\mu_{22}}{\mu_{21}\mu_{12}}$$

Intervalle de confiance de niveau $1-\alpha$:

$$\left[\widehat{RC}\ e^{-z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC}))},\ \widehat{RC}\ e^{z_{\alpha/2}\hat{\sigma}(\ln(\widehat{RC}))}\right] \quad \text{avec } \hat{\sigma}(\ln(\widehat{RC})) = \sqrt{\frac{1}{\mu_{11}} + \frac{1}{\mu_{12}} + \frac{1}{\mu_{21}} + \frac{1}{\mu_{22}}}$$

2.4 Cas particulier des variables ordinales

Coefficient de corrélation :
mesure d'association linéaire (Pearson) ou monotone (Spearman) entre X et Y

$$r = \frac{\displaystyle\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}vm_i wm_j - \frac{1}{n}\left(\sum_{i=1}^{I}\mu_{i\bullet}vm_i\right)\left(\sum_{j=1}^{J}\mu_{\bullet j}wm_j\right)}{\sqrt{\left(\displaystyle\sum_{i=1}^{I}\mu_{i\bullet}vm_i^2 - \frac{1}{n}\left(\sum_{i=1}^{I}\mu_{i\bullet}vm_i\right)^2\right)\left(\sum_{j=1}^{J}\mu_{\bullet j}wm_i^2 - \frac{1}{n}\left(\sum_{j=1}^{J}\mu_{\bullet j}wm_j\right)^2\right)}}$$

où les vm_i et wm_j sont des quantités représentant les modalités des variables :

– pour le **coefficient de Pearson** (r_P), ces modalités modifiées sont des scores numériques subjectifs mais choisis de façon à être les plus représentatifs possibles de la réalité ;
– pour le **coefficient de Spearman**, ces modalités modifiées sont les rangs moyens des modalités de X et Y, que l'on peut calculer ainsi :
rang moyen de la modalité v_i de X = $\frac{(\text{rang min} + \text{rang max})}{2}$
avec rang min = (nombre d'observations pour lesquelles $X < v_i$) + 1
rang max = (nombre d'observations pour lesquelles $X \leq v_i$)
Le même raisonnement s'applique aux rangs moyens des observations de Y.

Test d'association entre X et Y

$$H_0 \ : X \text{ et } Y \text{ ne sont sont pas associées}$$
$$H_1 \ \begin{cases} X \text{ et } Y \text{ sont associées (test bilatéral)} \\ X \text{ et } Y \text{ sont associées positivement} \\ X \text{ et } Y \text{ sont associées négativement} \end{cases}$$

Statistique de test :

$$M = r\sqrt{(n-1)} \ \xrightarrow[H_0]{\text{asymp.}} \ N(0,1)$$

ou encore la statistique de Mantel et Haenszel (test bilatéral uniquement) :

$$M^2 = (n-1)r^2 \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_1$$

où r est, au choix, la corrélation de Pearson (ass. linéaire) ou de Spearman (ass. monotone).

2.5 Cas particulier des données pairées

Sensibilité et spécificité d'un examen diagnostic :
X = Résultat du test diagnostique, soit v_1=positif (malade) ou v_2 = négatif (sain)
Y = Vrai état d'une personne, soit w_1=malade ou w_2 = sain
– **sensibilité** = $P(X = \text{positif} \mid Y = \text{malade})$
– **spécificité** = $P(X = \text{négatif} \mid Y = \text{sain})$

Test de la symétrie de la loi conjointe dans un tableau $I \times I$:

$$\begin{array}{lcll} H_0 & : & \pi_{ij} = \pi_{ji} & \text{pour tout couple } (i,j) \\ H_1 & : & \pi_{ij} \neq \pi_{ji} & \text{pour au moins un couple } (i,j) \end{array}$$

– **Statistique du khi-deux de Pearson (test de Bowker)** :

$$X^2_{sym} = \sum_{1 \leq i < j \leq I} \frac{(\mu_{ij} - \mu_{ji})^2}{\mu_{ij} + \mu_{ji}} \qquad \xrightarrow[H_0]{\text{asymp.}} \qquad \chi^2_{\frac{I(I-1)}{2}}$$

– **Statistique du rapport de vraisemblance** :

$$G^2_{sym} = \sum_{i=1}^{I}\sum_{j=1}^{I} \mu_{ij} \ln \frac{2\mu_{ij}}{\mu_{ij}+\mu_{ji}} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_{\frac{I(I-1)}{2}}$$

Test d'homogénéité des marginales dans un tableau $I \times I$:

$$\begin{array}{lcll} H_0 & : & \pi_{i.} = \pi_{.i} & \text{pour tout } i = 1, \ldots, I \\ H_1 & : & \pi_{i.} \neq \pi_{.i} & \text{pour au moins un } i \end{array}$$

On connait l'existence des statistiques de Stuart-Maxwell et de Bhapkar et on sait comment les calculer en SAS, mais leurs formules n'ont pas été données en classe. Sous H_0, ces statisiques suivent asymptotiquement une χ^2_{I-1}.

Test de McNemar :
test de symétrie de la loi conjointe et d'homogénéité des marginales pour un tableau 2×2 :

$$X^2_{sym} = \frac{(\mu_{12}-\mu_{21})^2}{\mu_{12}+\mu_{21}} \quad \xrightarrow[H_0]{\text{asymp.}} \quad \chi^2_1$$

Mesures d'accord entre les variables X et Y :

– **Proportion d'accord observé** :
Définition théorique :

$$P_0 = \sum_{i=1}^{I} \pi_{ii}$$

Estimateur :

$$\hat{P}_0 = \sum_{i=1}^{I} \hat{\pi}_{ii} = \sum_{i=1}^{I} \mu_{ii}/n$$

– **Kappa de Cohen** :
Définition théorique :

$$\kappa = \frac{P_0 - P_e}{1 - P_e}$$

où $P_e = \sum_{i=1}^{I} \pi_{i\bullet}\pi_{\bullet j}$ est la proportion d'accord aléatoire
Estimateur :

$$\kappa = \frac{\hat{P}_0 - \hat{P}_e}{1 - \hat{P}_e} \qquad \text{où} \qquad \hat{P}_e = \sum_{i=1}^{I} \frac{\mu_{i\bullet}}{n}\frac{\mu_{\bullet j}}{n}$$