1 Chapitre 5

1.1 lois Bernoulli et binomiale

- Une épreuve de Bernoulli(p) consiste à observer un succès ou un échec selon les probabilités

$$\begin{aligned} P(X=0) &= 1-p=q \\ P(X=1) &= p, \end{aligned}$$

où $0 \leq p \leq 1$.

- La moyenne et la variance sont

$$\begin{aligned} E(X) &= p \\ Var(X) &= p(1-p). \end{aligned}$$

- La fonction génératrice des moments (fgm) est

$$\phi(t) = q + pe^t.$$

- On obtient une variable binomiale(n, p) lorsqu'on observe le nombre de succès dans une suite de n épreuves de Bernoulli indépendantes. La fonction de masse est

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}, \ i = 0, 1, \ldots, n.$$

- Une variable binomiale s'écrit comme

$$X = \sum_{i=1}^{n} X_i,$$

où les variables X_i sont iid de loi Bernoulli(p). Ceci donne la moyenne et la variance

$$\begin{aligned} E(X) &= np \\ Var(X) &= np(1-p) \end{aligned}$$

de même que la fgm

$$\phi(t) = (q + pe^t)^n.$$

- La formule récursive est aussi utile

$$P(X=k+1) = \frac{p}{1-p}\frac{n-k}{k+1}P(X=k).$$

1.2 loi de Poisson

- Une variable X suit la loi de Poisson(λ), $\lambda > 0$, si X a pour fonction de masse

$$P(X = i) = e^{-\lambda}\frac{\lambda^i}{i!},\ i = 0, 1, \ldots$$

- La fgm de cette loi est

$$\phi(t) = \exp[\lambda(e^t - 1)]$$

ce qui procure la moyenne et la variance

$$\begin{aligned} E(X) &= \lambda \\ Var(X) &= \lambda. \end{aligned}$$

- La loi binomiale(n, p), où n est grand et p est petit peut être approchée par une loi de Poisson(λ), où $\lambda = np$.

- Si X_1 de loi Poisson(λ_1) est indépendante de X_2 de loi Poisson(λ_2) alors, $X_1 + X_2$ est de loi Poisson($\lambda_1 + \lambda_2$). C'est cette propriété qu'on invoque pour dire que si le nombre de particules alpha émises dans un intervalle d'une seconde suit une loi de Poisson$(3, 2)$ alors, le nombre de particules alpha émises dans un intervalle de 2 secondes suit une loi de Poisson$(6, 4)$.

- Si X_1 de loi Poisson(λ_1) est indépendante de X_2 de loi Poisson(λ_2) alors, la loi conditionnelle de X_1, étant donnée que $X_1 + X_2 = n$, est la loi binomiale$(n, \lambda_1/(\lambda_1 + \lambda_2))$.

- On a aussi la formule récursive

$$P(X = i + 1) = \frac{\lambda}{i + 1}P(X = i).$$

1.3 loi hypergéométrique

- Une urne contient N boules rouges et M boules noires. On choisit au hasard (sans remise) n boules de l'urne. La fonction de masse de X, le nombre de boules rouges sélectionnées, est celle de la loi hypergéométrique(N, M, n)

$$P(X = i) = \frac{\binom{N}{i}\binom{M}{n-i}}{\binom{N+M}{n}},\ \max(0, n - M) \leq i \leq \min(N, n).$$

- La moyenne de X est

$$E(X) = \frac{nN}{(N + M).}$$

- Si la sélection était faite avec remise, la loi de X serait binomiale$(n, N/(N + M))$.

- Si X de loi binomiale(n, p) est indépendante de Y de loi binomiale(m, p) alors, la loi conditionnelle de X, étant donnée que $X + Y = k$, est la loi hypergéométrique(n, m, k).

1.4 loi uniforme

- La variable X de loi uniforme(α, β) a pour densité

$$f(x) = \frac{1}{\beta - \alpha},\ \alpha < x < \beta.$$

- Si U est de loi uniforme$(0, 1)$ alors, $X = \alpha + (\beta - \alpha)U$ est de loi uniforme(α, β).

- La moyenne et la variance sont

$$\begin{aligned} E(X) &= \frac{\alpha + \beta}{2} \\ Var(X) &= \frac{(\beta - \alpha)^2}{12}. \end{aligned}$$

1.5 loi normale

- La variable X de loi $N(\mu, \sigma^2)$ a pour densité

$$f(x) = \frac{1}{\sqrt{2\pi}\ \sigma} e^{-(x-\mu)^2/2\sigma^2}.$$

- Si Z est de loi $N(0, 1)$ alors, $X = \mu + \sigma Z$ est de loi $N(\mu, \sigma^2)$. Réciproquement, si X est de loi $N(\mu, \sigma^2)$ alors, $Z = (X - \mu)/\sigma$ est de loi $N(0, 1)$.

- Les tables de la loi $N(0, 1)$ donne la fonction de répartition

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

  Par symmétrie, $\Phi(-z) = 1 - \Phi(z)$.

- Si X est de loi $N(\mu, \sigma^2)$ alors,

$$P(a < X < b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

- La fgm de la loi $N(\mu, \sigma^2)$ est

$$\phi(t) = e^{t\mu + \sigma^2 t^2/2.}$$

- Si X_i, $i = 1, \ldots, n$, sont indépendantes de loi $N(\mu_i, \sigma_i^2)$ alors $\Sigma_{i=1}^n X_i$ est de loi $N(\mu, \sigma^2)$ où $\mu = \Sigma_{i=1}^n \mu_i$ et $\sigma^2 = \Sigma_{i=1}^n \sigma_i^2$.

- Le quantile $1 - \alpha$ de la loi $N(0, 1)$ est noté z_α et satisfait

$$P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha.$$

1.6 loi exponentielle

- La variable X de loi exponentielle(λ), $\lambda > 0$, a pour densité

$$f(x) = \lambda e^{-\lambda x},\ x > 0.$$

- La fonction de répartition de X est $F(x) = 1 - e^{-\lambda x}$, $x > 0$.
- Si U est de loi exponentielle(1) alors, $X = U/\lambda$ est de loi exponentielle(λ).
- La fgm de X est $\phi(t) = \lambda/(\lambda - t)$, $t < \lambda$.
- La moyenne et la variance de X sont

$$\begin{aligned} E(X) &= 1/\lambda \\ Var(X) &= 1/\lambda^2. \end{aligned}$$

- La variable X est dite sans mémoire puisque

$$P(X > s + t | X > t) = P(X > s),\ s, t > 0.$$

- Si X_i, $i = 1, \ldots, n$, sont indépendantes de loi exponentielle(λ_i) alors, $\min(X_1, \ldots, X_n)$ est de loi exponentielle(λ), où $\lambda = \Sigma_{i=1}^n \lambda_i$. Cette propriété permet de modéliser la durée de vie d'un système de composantes indépendantes installées en série.

1.7 loi du khi-deux

- Si $Z_1, \ldots, Z_n$ sont iid $N(0, 1)$ alors, par définition, la loi de

$$X = Z_1^2 + \cdots + Z_n^2$$

  est la loi du χ_n^2.

- Le quantile $1 - \alpha$ de la loi du χ_n^2 est noté $\chi_{\alpha,n}^2$ et satisfait

$$P(X > \chi_{\alpha,n}^2) = \alpha.$$

- la fgm de X est $\phi(t) = (1 - 2t)^{-n/2}$, $t < 1/2$.
- Si X de loi χ_n^2 est indépendante de Y de loi χ_m^2 alors, $X + Y$ est de loi χ_{n+m}^2.
- La moyenne et la variance de la loi du χ_n^2 sont

$$\begin{aligned} E(X) &= n \\ Var(X) &= 2n. \end{aligned}$$

1.8 la loi de Student

- Si Z de loi $N(0,1)$ est indépendante de X_n de loi χ^2_n alors, par définition, la loi de

$$T_n = \frac{Z}{\sqrt{X_n/n}}$$

  est la loi t_n de Student à n degrés de liberté.

- La loi de T_n est symétrique par rapport à 0.

- Lorsque $n \to \infty$, la loi t_n converge vers la loi $N(0,1)$.

- La moyenne et la variance sont

$$\begin{aligned} E(T_n) &= 0 \\ Var(T_n) &= n/(n-2),\ n > 2. \end{aligned}$$

- Le quantile $1-\alpha$ de la loi t_n est noté $t_{\alpha,n}$ et satisfait

$$P(T_n > t_{\alpha,n}) = \alpha.$$

  Par symétrie, $-t_{\alpha,n} = t_{1-\alpha,n}$.

1.9 la loi F

- Si X_n de loi χ^2_n est indépendante de X_m de loi χ^2_m alors, par définition, la loi de

$$F_{n,m} = \frac{X_n/n}{X_m/m}$$

  est la loi F avec n et m degrés de liberté.

- Le quantile $1-\alpha$ de la loi $F_{n,m}$ est noté $F_{\alpha,n,m}$ et satisfait

$$P(F_{n,m} > F_{\alpha,n,m}) = \alpha.$$

- Si $F_{n,m}$ est de loi F à n et m degrés de liberté alors, $1/F_{n,m}$ est de loi F à m et n degrés de liberté. Cette propriété se traduit en terme des quantiles par

$$\frac{1}{F_{\alpha,n,m}} = F_{1-\alpha,m,n}.$$

2 Chapitre 6

Un échantillon (aléatoire) de taille n est représenté par des variables aléatoires $X_1, \ldots, X_n$ iid d'une certaine distribution de probabilité.

2.1 la moyenne

Pour un échantillon de taille n d'une distribution quelconque de moyenne μ et de variance σ^2, la moyenne échantillonale $\bar{X}$ a comme moyenne et variance

$$\begin{aligned} E(\bar{X}) &= \mu \\ Var(\bar{X}) &= \sigma^2/n. \end{aligned}$$

2.2 le TLC

- Pour un échantillon de taille n d'une distribution quelconque de moyenne μ et de variance σ^2,

$$\begin{aligned} P\left(\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \leq x\right) &\rightarrow \Phi(x),\ n \rightarrow \infty \\ P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq x\right) &\rightarrow \Phi(x),\ n \rightarrow \infty. \end{aligned}$$

- Si X est de loi binomiale(n, p) alors

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq x\right) \rightarrow \Phi(x),\ n \rightarrow \infty.$$

2.3 la variance

La variance échantillonale est définie par

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

et $S = \sqrt{S^2}$ est l'écart type. Pour un échantillon de taille n d'une distribution quelconque de moyenne μ et de variance σ^2, $E(S^2) = \sigma^2$.

2.4 échantillon d'une distribution normale

Pour un échantillon de taille n d'une distribution $N(\mu, \sigma^2)$,

- $\bar{X}$ est de loi $N(\mu, \sigma^2/n)$,
- $(n-1)S^2/\sigma^2$ est de loi χ^2_{n-1},
- $\bar{X}$ et S^2 sont des variables indépendantes,
- $\sqrt{n}(\bar{X} - \mu)/S$ est de loi t_{n-1}.

3 Chapitre 7

3.1 estimateur de vraisemblance maximale

On dispose d'un échantillon de taille n d'une distribution dont la densité (ou la fonction de masse) est $f(x|\theta)$. L'EVM est obtenu en optimisant la vraisemblance

$$L(\theta) = f(x_1|\theta) \dots f(x_n|\theta)$$

ou parfois la log-vraisemblance

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

3.2 intervalle de confiance sur la moyenne

- $X_1, \dots, X_n$ iid $N(\mu, \sigma_0^2)$, variance connue.

$$IC(\mu) : \bar{X} \pm z_{\alpha/2}\sigma_0/\sqrt{n}$$

$$IC(\mu) : \left(-\infty, \bar{X} + z_\alpha \sigma_0/\sqrt{n}\right)$$
$$IC(\mu) : \left(\bar{X} - z_\alpha \sigma_0/\sqrt{n}, +\infty\right)$$

- $X_1, \dots, X_n$ iid $N(\mu, \sigma^2)$, variance inconnue.

$$IC(\mu) : \bar{X} \pm t_{\alpha/2,n-1} S/\sqrt{n}$$

$$IC(\mu) : \left(-\infty, \bar{X} + t_{\alpha,n-1} S/\sqrt{n}\right)$$
$$IC(\mu) : \left(\bar{X} - t_{\alpha,n-1} S/\sqrt{n}, +\infty\right)$$

- $X_1, \dots, X_n$ (n grand) iid d'une distribution de moyenne μ et de variance σ^2 inconnue.

$$IC(\mu) : \bar{X} \pm z_{\alpha/2} S/\sqrt{n}$$

$$IC(\mu) : \left(-\infty, \bar{X} + z_\alpha S/\sqrt{n}\right)$$
$$IC(\mu) : \left(\bar{X} - z_\alpha S/\sqrt{n}, +\infty\right)$$

3.3 intervalle de confiance sur la variance

- $X_1, \dots, X_n$ iid $N(\mu, \sigma^2)$.

$$IC(\sigma^2) : \left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right)$$

$$IC(\sigma^2) : \left(\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}}, +\infty\right)$$

$$IC(\sigma^2) : \left(0, \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}\right)$$

3.4 Comparaison de deux moyennes, échantillons indépendants

- $X_1, \ldots, X_n$ iid $N(\mu_1, \sigma_1^2)$ indépendant de $Y_1, \ldots, Y_m$ iid $N(\mu_2, \sigma_2^2)$, les variances σ_1^2 et σ_2^2 connues.

$$IC(\mu_1 - \mu_2) : \bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$IC(\mu_1 - \mu_2) : \left(-\infty, \bar{X} - \bar{Y} + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$IC(\mu_1 - \mu_2) : \left(\bar{X} - \bar{Y} - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, +\infty\right)$$

- $X_1, \ldots, X_n$ iid $N(\mu_1, \sigma^2)$ indépendant de $Y_1, \ldots, Y_m$ iid $N(\mu_2, \sigma^2)$, la variance σ^2 inconnue. On estime la variance commune avec

$$S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}.$$

$$IC(\mu_1 - \mu_2) : \bar{X} - \bar{Y} \pm t_{\alpha/2, n+m-2}\ S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$IC(\mu_1 - \mu_2) : \left(-\infty, \bar{X} - \bar{Y} + t_{\alpha, n+m-2}\ S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

$$IC(\mu_1 - \mu_2) : \left(\bar{X} - \bar{Y} - t_{\alpha, n+m-2}\ S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, +\infty\right)$$

3.5 Comparaison de deux moyennes, mesures répétées

- $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid d'une distribution bivariée telle que les différences $D_i = X_i - Y_i$ suivent la loi $N(\mu, \sigma^2)$.

  On calcule

$$S_d^2 = \frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n-1}.$$

$$IC(\mu) : \bar{D} \pm t_{\alpha/2, n-1}\ S_d/\sqrt{n}$$

$$IC(\mu) : \left(-\infty, \bar{D} + t_{\alpha, n-1}\ S_d/\sqrt{n}\right)$$

$$IC(\mu) : \left(\bar{D} - t_{\alpha, n-1}\ S_d/\sqrt{n}, +\infty\right)$$

3.6 Intervalle de confiance pour une proportion

- $X_1, \ldots, X_n$ (n grand) iid Bernoulli(p).

  L'estimateur de p est $\hat{p} = \bar{X}$ et $\sum_{i=1}^n X_i$ est de loi binomiale(n, p). En utilisant le TLC pour la binomiale,

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0, 1).$$

$$IC(p) : \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

On peut déterminer la taille n de sorte que la longueur de l'intervalle est b pour un niveau donné $1-\alpha$ comme suit:

$$n = \frac{(2z_{\alpha/2})^2}{b^2}p(1-p)$$

ou

$$n \leq \frac{(z_{\alpha/2})^2}{b^2},$$

selon que l'on dispose ou non d'un estimé préliminaire de p.

3.7 Intervalle de confiance pour la moyenne d'une distribution exponentielle

- $X_1, \ldots, X_n$ iid exponentielle$(1/\theta)$, de moyenne θ.

  Dans ce cas on utilise la distribution

$$\frac{2}{\theta}\sum_{i=1}^{n} X_i \sim \chi^2_{2n}.$$

$$IC(\theta) : \left(\frac{2\sum_{i=1}^n X_i}{\chi^2_{\alpha/2,2n}}, \frac{2\sum_{i=1}^n X_i}{\chi^2_{1-\alpha/2,2n}}\right)$$

4 Chapitre 9

4.1 régression linéaire simple

Le modèle de régression linéaire simple est

$$y_i = \alpha + \beta x_i + e_i,\ i = 1, \ldots, n,$$

où les termes d'erreur $e_1, \ldots, e_n$ sont iid $N(0,1)$.

4.2 estimation de α et β par les moindres carrés

La méthode des moindres consiste à estimer α, l'ordonnée à l'origine, et β, la pente, au moyen du critère

$$\min_{A,B} \sum_{i=1}^{n} (y_i - A - Bx_i)^2.$$

La solution de ce problème est

$$\begin{aligned} B &= S_{xy}/S_{xx} \\ A &= \bar{y} - B\bar{x}, \end{aligned}$$

où

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ S_{xy} &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}. \end{aligned}$$

4.3 estimation de la variance σ^2

On aura besoin de

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

On estime ensuite la variance σ^2 au moyen des résidus $\hat{e}_i = y_i - A - Bx_i$,

$$\hat{\sigma}^2 = \frac{SS_R}{n-2},$$

où $SS_R = \sum_{i=1}^{n} \hat{e}_i^2$ est la somme de carrés résiduelle. On évalue SS_R par la formule

$$SS_R = S_{yy} - B^2 S_{xx}$$

sans avoir à calculer tous les résidus.

4.4 estimation par maximum de vraisemblance

Les estimateurs de α et β sont les mêmes que ceux des moindres carrés, à savoir A et B. L'estimateur de σ^2 diffère un peu

$$\tilde{\sigma}^2 = \frac{SS_R}{n}.$$

En pratique, on utilise plutôt l'estimateur sans biais $\hat{\sigma}^2$.

4.5 distributions d'échantillonage

$$\begin{aligned} B &\sim N\left(B, \frac{\sigma^2}{S_{xx}}\right) \\ A &\sim N\left(A, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right) \\ (n-2)\hat{\sigma}^2/\sigma^2 &\sim \chi^2_{n-2} \end{aligned}$$

Le vecteur aléatoire (A, B) est indépendant de $\hat{\sigma}^2$. Cependant, A et B ne sont généralement pas des variables indépendantes, sauf dans le cas où $\bar{x} = 0$.

4.6 intervalles de confiance

$$\begin{aligned} IC(\alpha) &: A \pm t_{\alpha/2,n-2}\ \hat{\sigma}\ \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{nS_{xx}}} \\ IC(\beta) &: B \pm t_{\alpha/2,n-2}\ \hat{\sigma}\ \frac{1}{\sqrt{S_{xx}}} \\ IC(\sigma^2) &: \left(\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2,n-2}}, \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2,n-2}} \right) \end{aligned}$$

4.7 intervalle de confiance sur la moyenne $E(y|x_0) = \alpha + \beta x_0$

Pour estimer la moyenne de la réponse y à une valeur donnée $x = x_0$ de la variable explicative on utilise la valeur sur la droite ajustée, à savoir

$$A + Bx_0$$

dont la distribution est

$$A + Bx_0 \sim N\left(\alpha + \beta x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

ce qui donne également l'intervalle de confiance

$$IC(\alpha + \beta x_0) : A + Bx_0 \pm t_{\alpha/2,n-2}\ \hat{\sigma}\ \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]^{1/2}.$$

4.8 intervalle de prévision d'une observation future $y_0 = \alpha + \beta x_0 + e_0$ à $x = x_0$

$$IP(y_0) : A + Bx_0 \pm t_{\alpha/2,n-2}\ \hat{\sigma}\ \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]^{1/2}.$$

4.9 Le coefficient de détermination R^2

Le coefficient de détermination R^2 est un indice de la qualité de l'ajustement donné par

$$R^2 = 1 - \frac{SS_R}{S_{yy}}$$

satisfaisant $0 \leq R^2 \leq 1$. L'ajustement est d'autant meilleur que la somme de carrés résiduelles SS_R est petite. Un bon ajustement correspond donc à des valeurs de R^2 proches de 1. On peut aussi montrer qu'en fait R^2 est le carré du coefficient de corrélation entre les x_i et les y_i. C'est aussi le carré du coefficient de corrélation entre les y_i et les valeurs prévues par le modèle, c'est-à-dire $\hat{y}_i = A + Bx_i$.