STT 1000 - STATISTIQUES

Arthur Charpentier

Moyenne et variance

Étant donné un échantillon $\{x_1, \cdots, x_n\}$, on appelle moyenne

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et on appelle variance empirique

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2,$$

Considérons maintenant une collection de variables aléatoires indépendantes et identiquement distribuées, $X_1, \cdots, X_n$, et définissons les variables aléatoires

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

Moyenne et variance

Si les variables X_i sont d'espérance μ et de variance σ^2,

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{n}{n}\mu = \mu$$

$$\text{Var}(\overline{X}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) = \frac{n}{n^2}\sigma^2 = \frac{\sigma^2}{n}$$

Moyenne empirique

Si les variables X_i sont indépendantes d'espérance μ et de variance σ^2,

$$\mathbb{E}(\overline{X}) = \mu \text{ et } \text{Var}(\overline{X}) = \frac{\sigma^2}{n}$$

Moyenne et variance

Si les variables X_i suivent des lois normales $\mathcal{N}(\mu, \sigma^2)$,

$$Z = n\frac{\overline{X} - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

Note si on suppose juste que les variables X_i sont d'espérance μ et de variance σ^2, le théorème central limite garantie que

$$Z_n = n\frac{\overline{X} - \mu}{\sigma} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

Moyenne et variance

$$\mathbb{E}\left(S^2\right)=\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2\right]=\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left([X_k-\mu]+[\mu-\bar{X}]\right)^2\right]$$

$$\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2=\sum_{k=1}^{n}[X_k-\mu]^2+\sum_{k=1}^{n}2[X_k-\mu][\mu-\bar{X}]+\sum_{k=1}^{n}[\mu-\bar{X}]^2$$

$$\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2=\underbrace{\sum_{k=1}^{n}[X_k-\mu]^2}_{n\sigma^2}+2[\mu-\bar{X}]\underbrace{\sum_{k=1}^{n}[X_k-\mu]}_{=n(\overline{X}-\mu)}+n[\mu-\bar{X}]^2$$

$$\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2=n\sigma^2-n[\mu-\bar{X}]^2$$

donc, en prenant l'espérance

$$\mathbb{E}\left(S^2\right)=\frac{1}{n}\left(n\sigma^2-n\frac{\sigma^2}{n}\right)=\frac{n-1}{n}\sigma^2$$

Moyenne et variance

Note: si les variables sont Gaussiennes $X_i \sim \mathcal{N}(\mu, \sigma^2)$,

$$\text{Var}[S^2] \sim \frac{2\sigma^4}{n}$$

Variance empirique

Si les variables X_i sont indépendantes sont d'espérance μ et de variance σ^2,

$$\mathbb{E}\left(S^2\right) = \frac{1}{n-1}\sigma^2 \sim \sigma^2 \text{ et } \text{Var}[S^2] \sim \frac{2\sigma^4}{n}$$

si $X_i \sim \mathcal{N}(0, \sigma^2)$ pour la variance.

Note: si les variables ne sont pas Gaussiennes

$$\text{Var}[S^2] = \frac{\mathbb{E}[(X-\mu)^4]}{n} - \frac{\sigma^4\,(n-3)}{n\,(n-1)}.$$

Médiane

Médiane empirique

Si les variables X_i sont de médiane m, de variance σ^2, et de densité f

$$\mathbb{E}[\text{mediane}(\boldsymbol{x})] \sim m \text{ et } \text{Var}[\text{mediane}(\boldsymbol{x})] \sim \frac{1}{4nf(m)^2}$$

Dans le cas d'une loi normale

$$\text{Var}[\text{mediane}(\boldsymbol{x})] = \frac{\pi\sigma^2}{2n} \sim \frac{1.2533^2\sigma^2}{n} > \frac{\sigma^2}{n} = \text{Var}[\overline{x}]$$

Soit m la médiane (théorique) i.e. $F(m) = \int_{-\infty}^{m} f(x)dx$

Pour la preuve, supposons que l'on dispose de $n = 2k + 1$ observations, i.e. la médiane (empirique) est $M = X_{(k+1)}$ ($(k+1)$ième observation) dont la densité est g

Médiane

$$g(y) = \frac{(2k+1)!}{k!k!} F(x)^k f(y)[1-F(x)]^k \sim \frac{(2k+1)4^k}{\sqrt{\pi k}} F(x)^k f(y)[1-F(x)]^k$$

(k observations sont plus petites que y, et k sont plus grandes).

$$F(x) \sim F(m) + F'(m) \cdot (x-m) = \frac{1}{2} + f(m) \cdot (x-m)$$

$$g(y) \sim \frac{(2k+1)4^k}{\sqrt{\pi k}} \left[\frac{1}{2} + f(m) \cdot (x-m)\right]^k f(y) \left[\frac{1}{2} - f(m) \cdot (x-m)\right]^k$$

$$g(y) \sim \frac{(2k+1)4^k}{\sqrt{\pi k}} \left[\frac{1}{2^2} - f(m)^2 \cdot (x-m)^2\right]^k f(y)$$

$$g(y) \sim \frac{2k \cdot f(y)}{\sqrt{\pi k}} \Bigg[1 - \underbrace{\frac{4kf(m)^2 \cdot (x-m)^2}{k}}_{=u/k}\Bigg]^k \sim \frac{2k \cdot f(y)}{\sqrt{\pi k}} \exp[-u]$$

Médiane

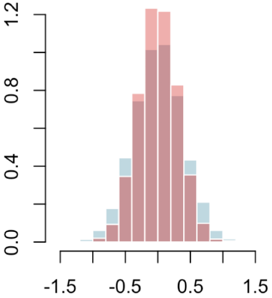
En posant $8kf(m)^2 = \gamma^{-1}$

$$g(y) \sim \frac{1}{\sqrt{\pi\gamma}} \exp\left[-\frac{-(y-m)^2}{2\gamma}\right]$$

Donc M suit, lorsque n est grand, une loi normale

$$M \sim \mathcal{N}\left(m, \frac{1}{8kf(m)^2}\right) \text{ ou } \mathcal{N}\left(m, \frac{1}{4nf(m)^2}\right)$$

```
> n = 11
> ns = 1e4
> U = matrix(runif(n*ns),ns,n)
> meanU = apply(U,1,mean)
> medianU = apply(U,1,median)
> var(meanU)
[1] 0.09216985
> var(medianU)
[1] 0.1407713
```



Échantillonnage en dimension finie

Supposons maintenant que les variables $X_1, \cdots, X_n$ sont tirées, sans remises, parmi N valeurs, $y_1, \cdots, y_N$

$$X_i = \begin{cases} y_1 \text{ avec probabilité } 1/N \\ y_2 \text{ avec probabilité } 1/N \\ \quad \vdots \\ y_N \text{ avec probabilité } 1/N \end{cases}$$

$$\mathbb{E}[X_j] = \sum_{i=1}^{N} \frac{1}{N} \mathbb{E}[X_j | X_j = y_i] = \frac{1}{N} \sum_{i=1}^{N} y_i = \mu = \overline{y}$$

Ici $\text{Cov}[X_j, X_k] \neq 0$. En effet

$$\mathbb{P}[X_j = y_u, X_k = y_v] = \underbrace{\mathbb{P}[Y_j = y_u]}_{1/N} \cdot \mathbb{P}[X_k = y_v | X_j = y_u]$$

$$\mathbb{P}[X_k = y_v | X_j = y_u] = \begin{cases} \dfrac{1}{N} \dfrac{1}{N-1} \text{ si } u \neq v \\ 0 \text{ si } u = v \end{cases}$$

Échantillonnage en dimension finie

$$\text{Cov}[X_j, X_k] = \mathbb{E}\big[(X_j - \mu)(X_k - \mu)\big] = \sum_{u,v=1}^{N} (y_u - \mu)(y_v - \mu)\mathbb{P}[X_j = y_u, X_k =$$

$$\text{Cov}[X_j, X_k] = \frac{1}{N(N-1)} \sum_{u \neq v = 1}^{N} (y_u - \mu)(y_v - \mu)$$

or

$$\left(\sum_{u=1}^{N} (y_u - \mu)\right)^2 = \sum_{u=1}^{N} (y_u - \mu)^2 + \sum_{u \neq v = 1}^{N} (y_u - \mu)(y_v - \mu)$$

donc $\text{Cov}[X_j, X_k] = -\dfrac{\sigma^2}{N-1}$

On peut alors calculer

$$\text{Var}[\overline{X}] = \frac{1}{n^2} \text{Var}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\left[\sum_{i=1}^{n} \text{Var}[X_i] + \sum_{j \neq k = 1}^{n} \text{Cov}[X_j, X_k]\right]$$

Échantillonnage en dimension finie

Moyenne dans un échantillon fini

Si les variables $X_1, \cdots, X_n$ sont tirées, sans remise, parmi N valeurs, $y_1, \cdots, y_N$,

$$\mathbb{E}[\overline{X}] = \mu \text{ et } \text{Var}[\overline{X}] = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

où $\mu = \frac{1}{N}\sum_{i=1}^{n} y_i$ et $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(y_i - \mu)^2$

Note: lorsque N est (très) grand, on retrouve $\text{Var}[\overline{X}] \sim \frac{\sigma^2}{n}$