Cours 04
Quantiles d'une loi normale
empirique, théorique, simulé

1 Quantiles, déciles $k/10$, fractiles k/d

Nous avons vu la définition des quartiles Q_1, Q_2, et Q_3 d'un échantillon ; c'est un cas particulier de quantile q d'une proportion $p = k/d$ (ou d-fractile) : on cherche un nombre q tel qu'une proportion p des individus de l'échantillon vérifient $x_i \leq q$. Lorsque $p = 10\%, 20\%, \ldots, 90\%$, les quantiles q_1, q_2, …, q_9 correspondants s'appellent des `déciles`. Ceci n'est généralement possible qu'approximativement (par exemple si la taille de l'échantillon n'est pas un multiple de d). En pratique, pour $p = k/d$ on choisira [1] $q_k =$`sort(x)[n*k/d]` , ou mieux : $q_k =$`sort(x)[ceiling(n*k/d)]`

2 Quantiles d'une loi normale

Désignons par $F_{\mu,\sigma}$ la fonction de répartition d'une loi normale $\mathcal{N}(\mu, \sigma)$ (ou gaussienne). En d'autres termes $F_{\mu,\sigma} = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2\sigma^2}} dt$; c'est par définition la *probabilité* d'être inférieure à x d'une grandeur aléatoire qui suit une loi $\mathcal{N}(\mu, \sigma)$. Ici il est facile de définir le quantile de $p_k = k/d$ de la loi $\mathcal{N}(\mu, \sigma)$: c'est le nombre [2] q_k tel que $F_{\mu,\sigma}(q_k) = p_k (= k/d)$. Il existe bien et est unique pourvu que $0 < p_k < 1$, puisque $F_{\mu,\sigma}$ est continue, strictement croissante, et de valeurs $F_{\mu,\sigma}(]-\infty, +\infty[) =]0, 1[$. Sa valeur est donnée, par `R`, par la fonction `qnorm(p,mu,sigma)`, si `p`$= p_k$, `mu`$= \mu$, et `sigma`$= \sigma$. En d'autres termes, c'est le nombre q_k tel que la probabilité d'être inférieure à ce nombre est égale à p_k. Nous voyons donc ici qu'on a juste remplacé "proportion" par "probabilité" en passant d'empirique [3] à théorique. En théorie des probabilités, c'est la Loi des Grands Nombres qui motive cette relation entre "proportion" et "probabilité".

3 Valeurs exceptionnelles d'un échantillon théorique

Considérons la loi gaussienne $\mathcal{N}(\mu, \sigma)$ associée à un échantillon `x`. La boite à moustaches théorique est alors donnée par $Q_1 =$`Q1=qnorm(0.25,mu,sigma)`$=$ 17.50790, $Q_2 =$`Q2=qnorm(0.50,mu,sigma)`$=$ 18.80238, $Q_3 =$`Q3=qnorm(0.75,mu,sigma)`$=$ 20.09687, et la longueur maximale des moustaches est $L =$`1.5*(Q3-Q1)`$=$ 3.883452.

Si `mu`$= \mu = 0$ et si `sigma`$= \sigma = 1$, l'interquartile vaut $Q_3 - Q_1 = 1.34898$ et la longueur maximale des moustaches vaut $1.5(Q_1 - Q_3) = 2.023469$. Les valeur exceptionnelles de `x` sont donc celles qui sont inférieures à $q_{min} = Q_1 - L = -2.697959$ ou supérieures à $q_{max} = Q_3 + L = 2.697959$. La probabilité p_{min} d'être inférieur à q_{min} est alors de 0.003488302, et la probabilité p_{max} d'être supérieur à q_{max} est alors de 0.003488302

4 Comparaison des quantiles empiriques et théorique

Nous avons déjà vu qu'en statistique un échantillon `x` présentant un histogramme en cloche suggère un modèle gaussien $\mathcal{N}(\mu, \sigma)$, avec $\mu =$`mean(x)` et $\sigma =$`sd(x)`, qu'on peut "tester" visuellement, en superposant à l'histogramme (en densité ou proportions e_k/n, par `freq=F`)) la courbe du graphe de la fonction de densité de $\mathcal{N}(\mu, \sigma)$. Une meilleure façon de tester si un modèle gaussien pour `x` est pertinent est de comparer les fractiles q_k de $p_k = k/d$ de l'échantillon et les fractiles théoriques $F^{-1}_{\mu,\sigma}(p_k)$, pour $k = 1, \ldots, d-1$. Pour l'échantillon `x=survey.cc$WrHnd`, nous trouvons $\mu =$`mean(x)`$=$ 18.80238 et $\sigma =$`sd(x)`$=$ 1.919205. Nous obtenons, pour $d = 4$, les quartiles empiriques et théoriques suivants :

empirique	17.5	18.5	20.0
théorique	17.50790	18.80238	20.09687

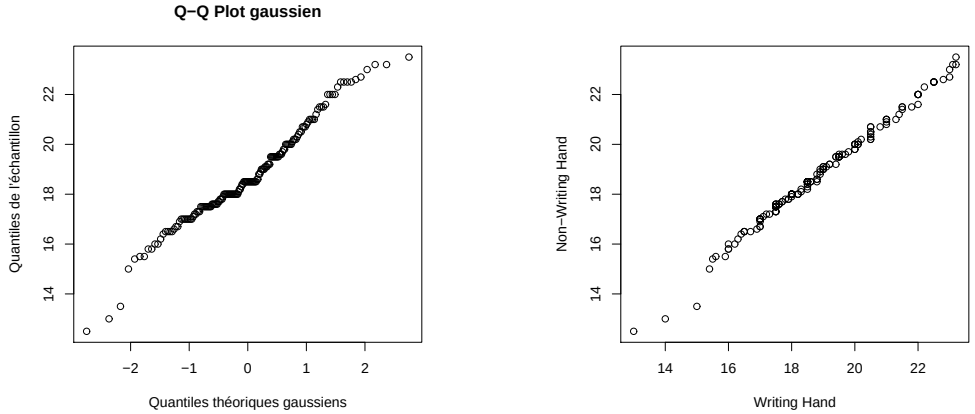
1. `R` accepte des numéros non entiers : il utilise alors la partie entière du nombre passé comme numéro.
2. ou encore $q_k = F^{-1}_{\mu,\sigma}(p_k)$, où $F^{-1}_{\mu,\sigma}$ *fonction réciproque* de $F_{\mu,\sigma}$, qui est définie pour tout $p \in]0, 1[$.
3. empirique : qui (ne) s'appuie (que) sur l'expérience ; nous utilisons ici le mot "empirique" pour désigner ce qui est relatif à un échantillon issue de la mesure d'un caractère pour une famille d'individus, comme ceux fournis par la bibliothèque (en anglais : library) `MASS` de `R`.

Pour $d = 10$, nous obtenons les déciles empiriques et théoriques suivants :

empirique	16.5	17.5	17.6	18.0	18.5	18.9	19.5	20.5	21.5
théorique	16.34282	17.18714	17.79595	18.31616	18.80238	19.28861	19.80881	20.41762	21.26194

5 `QQ-plot` : comparaison de quantiles

La commande `QQ-plot` permet de comparer les quantiles d'un échantillon avec ceux d'une loi normale (aussi appelée gaussienne), ou les quantiles de deux caractère distincts. Voici le résultat pour l'échantillon `NWHnd` de `survey`, comparé à une lois gaussienne centrée réduite, et à l'échantillon `WrHnd`.



Rappelons qu'on avait les boîtes à moustaches suivantes pour `WrHnd` (à gauche) et `NWHnd` (à droite). Les paliers observés sur le QQ-plot de `NWHnd` contre une loi gaussienne semblent dûs au fait que les décimales `0` et `5` semblent avoir été favorisées dans les mesures relevées pour ce caractère, comme le révèle l'histogramme des parties décimales de `NWHnd` (à droite, ci-dessous).