TP du CC2–QQ-plot (graphiques quantile-quantile)

Soit $x_1, \cdots, x_n$ une série statistique. On peut chercher à savoir si le distribution des données est gaussienne ou Poisson etc... Notons F_0 la fonction de répartition de cette loi de probabilitée d'intérêt. Le QQ-plot est un outil graphique permettant de visualiser rapidement l'adéquation de la distribution d'une séerie numérique à une distribution de référence. Dans ce graphe, on reporte sur l'axe des ordonnées les fractiles correspondant à la distribution observée et sur l'axe des abscisses ceux correspondant à la distribution théorique.

2.1 En pratique

Tableau des quantiles et nuage de points

- *Dans le cas d'une variable quantitative dont les valeurs sont regroupées par modalités* : Soient $m_1, \ldots, m_J$ les modalités de la série $x_1, \ldots, x_n$ que l'on appellera quantile observés ($q_i = m_i$). On remplit le tableau des fréquences cumulées. Pour chaque fréquence cumulée, on calcule le quantile théorique i.e. q_j^* tel que

$$F_0(q_j^*) = F_j \quad \Leftrightarrow \quad q_i^* = F_0^{-1}(F_j)$$

En général ce calcul se fait par l'utilisation des tables statistiques ou par un logiciel.
On reporte dans un graphique le nuage de points $(q_j, q_j^*)_{j=1\ldots J}$.

Modalités ordonnées, quantiles observés	$q_1 = m_1$	$\cdots$	$q_J = m_J$
Fréquences cumulées	F_1	$\cdots$	F_J
Quantiles théoriques	$q_1^* = F_0^{-1}(F_1)$	$\cdots$	$q_J^* = F_0^{-1}(F_J)$

Note : Les quantiles théoriques d'une loi normale $q_J^* = F_0^{-1}(F_J)$ sont accessibles sur R via `qnorm(`F_J, m, σ`)`.

2.2 Interprétation

- Si les points sont alignées sur la diagonale du carré de côté 1 (1ère bissectrice), alors la loi théorique proposée (de fonction de répartition F_0) est adaptée aux observations.
- Si les points sont alignés sur une droite parallèle à la diagonale du carré de côté 1 on soupçonnera une erreur sur les paramètres de position de la loi théorique.
- Si les points sont alignés sur une droite passant par l'origine mais inclinée par rapport à la diagonale du carré de côté 1 on soupçonnera une erreur sur les paramètres de dispersion de la loi théorique.
- Si les points sont alignés sur une droite ne passant pas par l'origine et inclinée par rapport à la diagonale du carré de côté 1 on soupçonnera une erreur sur les paramètres de dispersion et de position de la loi théorique.
- Si les points ne sont pas alignés sur une droite la loi théorique n'est pas adaptée aux observations.

Remarque
Si l'on dispose des données individuelles d'une variable aléatoire continue, les modalités sont toutes les valeurs prises par la série (ordonnées) et les fréquences cumulées sont du type $\frac{i}{n}$.

Remarque
Les QQ-plot peuvent servir aussi à comparer les distributions de deux séries.

2.3 Exercice

On considère la série de relevés de longueurs d'une série manufacturée de 500 pièces au dixième de millimiètre.

x_1	9.2	9.3	9.4	9.5	9.6	9.7	9.8	9.9	10	10.1	10.2	10.3	10.4	10.5	10.6	10.7	10.8	10.9
n_i	1	4	9	28	23	37	51	63	65	60	48	38	32	23	12	3	1	2

Calculer m et σ les moyenne et écart-type de la série x.

On souhaite savoir si cette série est proche d'une loi Gaussienne de moyenne m et décart-type σ. Tracr le QQ-plot et conclure.

3 Corrigé

- Saisie des q_i (noter que les q_i sont les modalités x_i) :

```
qi=c(9.2,9.3,9.4,9.5,9.6,9.7,9.8,9.9,10,10.1,10.2,10.3,10.4,10.5,10.6,10.7,10.8,10.9)
```

- Saisie des effectifs :

```
ni=c(1,4,9,28,23,37,51,63,65,60,48,38,32,23,12,3,1,2)
```

- Calcul de la moyenne pondérée :

```
m=sum(qi*ni)/sum(ni),
```

  On obtient $m = 10.01$.

- Calcul de l'écart-type pondéré :

```
sigma=sqrt(sum(qi*qi*ni)/sum(ni)-m^ 2),
```

  l'écart-type vaut $\sigma = 0.31$.

- **On souhaite donc observer l'adéquation avec une loi Normale $\mathcal{N}(m, sigma)$, avec m et σ vues dans les points précédents.**

- Calcul des fréquences cumulées :

```
Fi=cumsum(ni)/sum(ni).
```

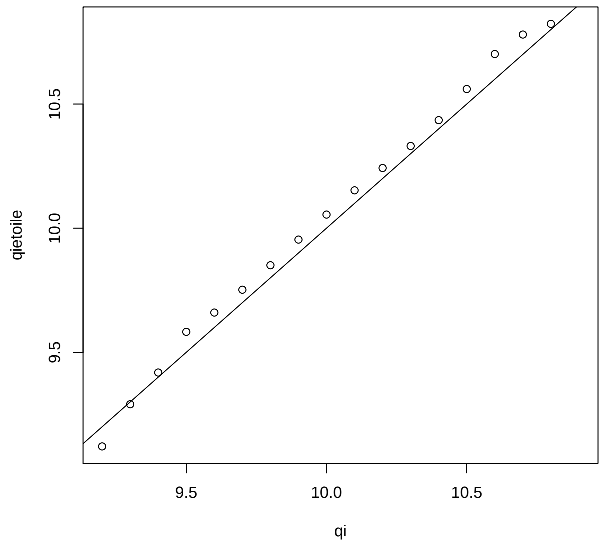
- Quantiles théoriques de la loi normale :

```
qietoile=qnorm(Fi,m,sigma)
```

- On trace le qqplot et on ajoute la diagonale :

```
plot(qi,qietoile)
lines(c(9,11),c(9,11))
```

- Interprétation



L'adéquation à la loi théorique choisie parraît bonne, la distribution semble gaussienne.