Cours 9 : Test du χ^2

Nous abordons une nouvelle démarche statistique inférentielle : celle de test d'hypothèse. Il s'agit d'une démarche de réfutation : on envisage de réfuter une hypothèse, comme par exemple qu'une pièce de monnaie donnée à les mêmes chances de tomber sur pile que sur face. Réfuter cette hypothèse conduit à conclure que cette pièce est pipée. L'hypothèse envisagée est notée H_0 et est appelée l'*hypothèse nulle*. Dans notre exemple l'hypothèse nulle est donc que la probabilité de pile est $p = 0.5$. Si on procède à n tirages, on n'est pas surpris de ne pas obtenir $n/2$ piles et autant de face, mais la question est de savoir à partir de quel écart à cette valeur théorique il convient de rejeter l'hypothèse nulle. Il y aura bien sûr un seuil de confiance α qui fixera la probabilité de rejeter à tort cette hypothèse. La méthode envisagée ici est une alternative facile à mettre en oeuvre à la simple vérification que les résultats observés ne sont pas dans un intervalle de confiance de la valeur théorique sous l'hypothèse H_0.

1 Mise en place de la statistique

Il convient de formuler l'hypothèse nulle H_0 de façon à ce qu'elle permette clairement, devant un échantillon, de déterminer les quantité suivantes :

- les classes numérotées généralement par $i = 1, \ldots, k$, où k désigne le nombre de classes,
- les nombres O_i d'individus de l'échantillon qui sont dans la classe i ; on les appelle les *effectifs observés*,
- les nombres T_i d'individus qui seraient idéalement dans la classe i sous l'hypothèse H_0. Souvent $T_i = np_i$ si H_0 fixe la probabilité d'être dans la classe i, où n est la taille de l'échantillon. On les appelle les *effectifs théoriques*.

Avec ces nombres on forme alors la *statistique*

$$S = \sum_{i=1}^{k} \frac{(O_i - T_i)^2}{T_i}.$$

On comprend facilement qu'on est d'autant plus enclin à rejeter l'hypothèse nulle que les effectifs observés s'écartent des effectifs théoriques, situation qui correspond à une statistique S élevée. Il convient alors de voir, en fonction de la valeur du seuil de confiance α, à partir de quelle valeur de la statistique S on peut rejeter l'hypothèse nulle.

2 La loi du χ^2 suivie par S

Dès que $k-1$ effectifs observés sont connus, l'effectif de la dernière classe est déterminé, par soustraction à l'effectif total n de l'échantillon. On dit que S a $k-1$ degrés de liberté. Une preuve probabiliste montre que, pour n assez grand, la loi de S peut être assimilée à une *loi du* $\chi^2(k-1)$, dite à $k-1$ degrés de liberté. Dans `R`, pour la loi du χ^2, les instructions usuelles pour une loi sont notées en utilisant le radical `chisq`, pour "chi squared". Ainsi la probabilité d'être inférieure à x pour une variable aléatoire S suivant une loi $\chi^2(k-1)$ est donnée par `pchisq(x,k-1)`. De façon analogue, le quantile q tel que la soit égale à p la probabilité que S soit inférieurs à q est donnée par l'instruction `qchisq(p,k-1)`. La densité de la loi du $\chi^2(k-1)$ au point x vaut `dchisq(x,k-1)`.

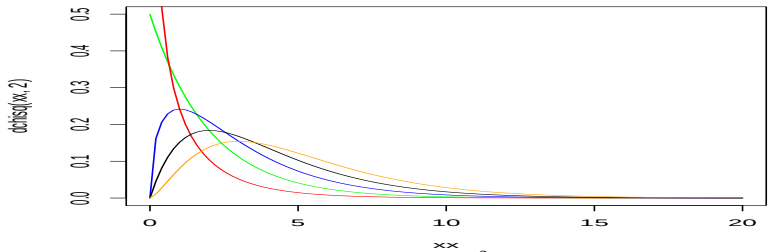


FIGURE 1 – Les densités de la loi du χ^2 à 1,...,5 degrés de liberté.

```
#        Cours de Stat Mass 09: dés, et Test du Chi2 d'un dé
######################## simulation d'un dé  ##############################
n=900
u= runif(n);u
dé=1+floor(6*u);dé
bornes=0.5:6.5;bornes
hist(dé,breaks=bornes,freq=TRUE)
# décompte des Observées (effectifs empiriques)
O=numeric(6)
for (j in 1:6)
{O[j]=sum(dé==j)}
O;sum(O)
# avec mon échantillon je trouve  139 148 151 156 149 157

#################  un dé pipé à présent ################################
# simulation d'un dé pipé
pipage=0.04; # pipage<1/6
z=sapply(1+6*(pipage+u),floor);z  # on décale: moins sur 1, plus sur 6
dé.pipé=numeric(n) #on crée un vecteur pour ranger les effectifs des classes de l'échantillon
for (i in 1:n)
{dé.pipé[i]=min(z[i],6)}  # et on met sur 6 ce qui irait sur 7
dé.pipé
# autre pipage (en réserve):
# dé.pipé=floor(1+6*(pipage+u*(1-pipage)));dé.pipé # on affaiblit 1 et on réparti
# sur les autres
hist(dé.pipé,breaks=bornes,freq=TRUE)
# hist(dé.pipé,breaks=bornes,include.lowest = FALSE, right = FALSE,freq=FALSE)
O.pipé=numeric(6)  #on crée un vecteur pour ranger les valeurs du test
for (j in 1:6) # on va envisager les 6 valeurs possibles
{O.pipé[j]=sum(dé.pipé==j)}  # nombre de tirages de la valeur j:
O.pipé;sum(O.pipé) # affichage des valeurs obtenues: on retrouve l'histogramme
# avec mon échantillon je trouve 111 136 149 156 146 202

################################ test du chi2  ###################################
# Rappel, on avais trouvé les effectifs observés suivants
O;sum(O)

# le test du Chi2: calcul de
S=sum((O-n/6)^2/(n/6));S   # la statistique S; je trouve T=1.413333

# il y a 6 faces donc 5 degrés de liberté
# au seuil alpha, T doit être inférieur à qchisq(1-alpha,5)
alpha=0.10        # seuil 10%
qchisq(1-alpha,5) # je trouve q= 9.236357 (quantile 1-alpha d'un loi du chi2(5) )
alpha=0.01        # seuil 5%
qchisq(1-alpha,5) # je trouve q= 15.08627
# comme ma statistisque du chi2 donne T=6.72,et qu'elle est inférieur à la p-value
# au seuil alpha=10%, on ne rejette pas l'hypothèse H_0 que le dé n'est pas pipé.

# calcul de la statistique du dé pipé
# Rappel: on avait trouvé:
O.pipé;sum(O.pipé)
# Calcul de la statistique:
S.pipé=sum((O.pipé-n/6)^2/(n/6));S.pipé   # je trouve la statistique S=30.78667
# comme ma statistisque du chi2 donne T=29.82667,et qu'elle est supérieurs au
# quantile 1-alpha de chi2(5)
# au seuil alpha=1%, on rejette l'hypothèse H_0 que le dé n'est pas pipé.
#                   "au seuil alpha=1% le dé est pipé"

1-pchisq(S.pipé,5) # je trouve la proba de 1.595356e-05 d'obtenir une si grande
                   # valeur ou plus
```