LES DISTRIBUTIONS DISCRETES ET CONTINUES

Veillez à avoir lu les sections relatives à cette séance dans le document disponible sur le site « Introduction aux séances de TP de biostatistique » avant de vous présenter au TP.

Veillez aussi à maitriser les notions apprises au(x) TP(s) précédent(s).

Préambule

Certaines fonctions statistiques permettent de calculer des probabilités de manière très simple.

Citons, parmi d'autres, les **fonctions EXCEL** suivantes:

ALEA() génère un nombre aléatoire compris entre 0 et 1 suivant une loi uniforme. **ALEA.ENTRE.BORNES(min ;max)** génère ce nombre entre une limite inférieure **min** et une limite supérieure **max**.Les valeurs générées par cette fonction sont toujours des valeurs entières.

LOI.BINOMIALE.N(r;n;p;cumulative) permet de calculer des probabilités selon la loi binomiale; **r** est le nombre de fois où l'événement d'intérêt a été observé, **n** est le nombre de fois qu'a été réalisée l'expérience, **p** est la probabilité que l'événement se réalise lorsqu'on effectue une fois l'expérience, et **cumulative** permet de calculer la probabilité pour la seule valeur de r (FAUX), ou pour toutes les valeurs de 0 à r (VRAI). Par exemple, la probabilité donnée par « =LOI.BINOMIALE.N(3;6;0,25;VRAI) » retourne la probabilité cumulée d'avoir entre 0 et 3 (inclus) réalisations d'un événement ayant une probabilité d'occurrence de 0,25 lors de chaque essai lorsqu'on effectue 6 essais (soit P(r<=3)).

LOI.HYPERGEOMETRIQUE.N(a1;N1;a;N;cumulative) permet de calculer une probabilité hypergéométrique dans le cas suivant: supposons qu'un échantillon de taille **N** puisse être découpé en deux sous échantillons (par exemple, les mâles et les femelles), de tailles respectives **N1** et N2=N-N1. Si **a** individus possèdent une propriété particulière dans l'échantillon (par exemple le fait d'être malades), la fonction permet alors de calculer la probabilité que **a1** individus, parmi ces **a**, appartiennent au sous-échantillon de taille **N1**. **Cumulative** permet de calculer la probabilité pour la seule valeur de a1 (FAUX), ou pour toutes les valeurs de 0 à a1 (VRAI).

LOI.KHIDEUX.N(X;ddl;cumulative) permet de calculer la (densité de) probabilité associée à une valeur **X** de χ^2 avec un nombre donné de degrés de liberté (**ddl**). Cumulative permet de dire si on désire la densité de probabilité associée à **X** (FAUX), ou la probabilité que χ^2 soit inférieur à X (VRAI). Par exemple, « =LOI.KHIDEUX.N(3,841;1 ;VRAI) » correspond à P(X < 3,841) pour 1 degré de liberté, ce qui retourne approximativement 0,95. Nous utiliserons VRAI dans la majorité des exercices !

LOI.NORMALE.N(X;μ;σ;cumulative) permet de calculer la (densité de) probabilité associée à une valeur **X** d'une loi normale de moyenne **μ** et de déviation standard **σ. Cumulative** permet de dire si on désire la densité de probabilité associée à **X** (FAUX), ou la probabilité d'être inférieur à X (VRAI). Par exemple, si on tape « =LOI.NORMALE.N(1,96;0;1;VRAI) », on calcule P(X < 1,96) pour une distribution normale de moyenne 0 et écart-type 1 (c'est-à-dire selon la loi normale standard). La valeur obtenue est (approximativement) 0,975. Nous utiliserons VRAI dans la majorité des exercices !

LOI.NORMALE.STANDARD.N(z ;cumulative) est équivalente à la précédente, mais ne s'applique que pour la loi normale standard (c'est-à-dire de moyenne $\mu = 0$ et de déviation standard $\sigma = 1$).

A noter qu'il existe les fonctions inverses LOI.BINOMIALE.INVERSE, LOI.NORMALE.INVERSE et KHIDEUX.INVERSE pour ces distributions : étant donné une probabilité et les paramètres de la fonction, on retrouve la valeur de la variable aléatoire.

Par exemple, la fonction « = LOI.NORMALE.INVERSE.N(0,975;0;1) » fournit à peu près la valeur de X = 1,96.

Dans R, on utilise 4 lettres selon le calcul souhaité (avec X le nom de la distribution utilisée, voir ci-dessous) :

- *dX()* : donne la densité de probabilité (distribution continue) ou la probabilité (distribution discrète) associée à la valeur fournie en argument (en plus des paramètres de la distribution utilisée) pour la distribution X.
- *pX()* : donne la probabilité d'être inférieur ou égal à la valeur fournie en argument (en plus des paramètres de la fonction utilisée) pour la distribution X. Pour obtenir la probabilité d'être supérieur à la valeur fournie, il faut ajouter l'argument *lower.tail=FALSE* à la fonction.
- *qX()* : est la fonction inverse de pX() : on fournit une probabilité **p** comme argument (en plus des paramètres de la fonction utilisée) et la fonction retourne la valeur **v** de la variable aléatoire **x** telle que P(**x** ≤ **v**) = **p**. Pour obtenir **v** tel que Pr(**x** > **v**), il faut ajouter l'argument *lower.tail=FALSE* à la fonction.
- *rX()* : renvoie des valeurs aléatoires de la variable aléatoire de distribution X. Cette fonction permet notamment de générer aléatoirement un échantillon selon des paramètres fixés.

Les noms des distributions (*X*) dans R sont notamment :

- *unif* pour la distribution uniforme
- *pois* pour la distribution de Poisson
- *binom* pour la distribution binomiale
- *hyper* pour la distribution hypergéométrique
- *norm* pour la distribution normale
- *chisq* pour la distribution de chi-carré

Par exemple, si on souhaite calculer P(r<=3) selon une loi binomiale caractérisée par n=6 tirages et p=0,25, on écrira *pbinom(q=3,size=6,prob=0.25).* Dans cette écriture, on a utilisé des « arguments nommés » dans pbinom : 'q' représente la valeur de r, 'size' la valeur de n et 'prob' la valeur de p. Il est possible de ne pas préciser ces arguments « nommés » tant que les valeurs renseignées respectent l'ordre attendu par le logiciel.

L'utilisation des fonctions préalablement citées dans R et excel est également détaillée dans le document « dias présentées au TP2 ». N'oubliez pas d'aussi vous référer aux aides des logiciels (en faisant par exemple* ?pbinom *dans R pour savoir comment calculer des probabilités cumulées selon la loi binomiale).

Exercice 1

Si X a une distribution normale, calculez la probabilité d'avoir une valeur de X comprise entre 20 et 30 si μ vaut 22 et σ = 4. Calculez d'abord les valeurs de z1 et z2 (les variables centrées réduites normales, voir formule…) et calculez la probabilité, dans Excel et dans R, en utilisant ce qui a été décrit dans le préambule.

Calculez cette même probabilité sans avoir recours au calcul de z1 et z2 (donc en utilisant une distribution normale caractérisée par μ 22 et σ = 4).

Exercice 2

Quelle est la probabilité, dans un échantillon de taille 20, d'avoir entre 4 et 10 animaux atteints (limites incluses) par une pathologie ayant une prévalence de 20% ? Utilisez la distribution binomiale (malade >< sain). Faites le calcul dans Excel et dans R.

Dans R, calculez et tracez la distribution binomiale théorique correspondant aux paramètres donnés. La variable binomiale r sera donc l'abscisse et la probabilité l'ordonnée. Le graphique sera du type diagramme à barres (*barplot(Y,names.arg=X)* ou *plot(X,Y,type="h")*).

Exercice 3

Est-il plus probable d'obtenir une valeur de 6 dans une distribution de χ^2 avec 4 degrés de liberté, ou une valeur de 10 dans une distribution de χ^2 avec 8 degrés de liberté ? Pour répondre à cette question, calculez les deux probabilités demandées avec Excel.

Dans R, établissez ensuite numériquement et graphiquement les distributions théoriques de chi-carré pour 4 et 8 degrés de libertés. Pour superposer un graphique à un autre créé avec la fonction *plot()*, vous pouvez notamment utiliser l'instruction graphique secondaire *lines()*.

Calculez ensuite la probabilité d'obtenir une valeur de chi-carré inférieure ou égale à 6 (pour 4 ddl) et à 10 (pour 8 ddl). Utilisez pour cela la fonction *pchisq()*) et tracez un trait vertical à hauteur de ces deux valeurs d'abscisse sur votre graphique à l'aide de la fonction *abline()*.

Exercice 4

Reconstituez la table ci-dessous dans Excel, ainsi que toutes les autres tables qu'on peut obtenir à partir de ces totaux marginaux en faisant varier le nombre de mâles atteints. Calculez ensuite la probabilité exacte de cette table et de chacune des autres tables créées. Que vaut la somme des probabilités de toutes ces tables ? Faites un histogramme des probabilités (en fonction du nombre de mâles atteints).

	Mâles	Femelles	
Atteints	3		8
Sains			
	7		12

Dans R, créez la même table à l'aide de la fonction ***matrix(X, nrow=2, byrow=TRUE, dimnames=list(c("Atteints", "Sains"),c("Males","Femelles")))*** dans laquelle **X** représente un vecteur avec les effectifs des cellules de la table ci-dessous, **nrow** indique le nombre de lignes du tableau, **byrow=T** indique que les données (contenues dans X) sont introduites dans la table ligne par ligne et **dimnames** est un argument permettant de nommer les lignes et colonnes du tableau (dans cet ordre).

Remarque : vous pouvez aussi réaliser le tableau sans le nommer (sans l'argument *dimnames*) puis nommer la table une fois créée avec *colnames(table) <- c("Males","Femelles")*et *rownames(table) <- c("Atteints", "Sains")*

Ensuite, la transformation en table de contingence peut se faire via la fonction *as.table()*.

Calculez la probabilité d'observer cette table sous l'hypothèse nulle (càd. dans la situation où il y a proportionnellement autant d'atteints chez les mâles que chez les femelles). Pour cela, utilisez la fonction *dhyper()*.

Exercice 5

Dans une clinique vétérinaire canine, il y a en moyenne 1 torsion d'estomac par mois. Durant le mois d'août, les vétérinaires ont eu 3 cas de torsion d'estomac. Calculez la probabilité de cette situation avec Excel.

Dans R, calculez la probabilité d'en avoir 3 ou plus au cours d'un mois.

Exercices facultatifs pour le brain storming

1. Dans R, créez le tableau de l'exercice 4 à l'aide d'une deuxième méthode : créez deux vecteurs de 12 places (chaque place du vecteur représente un individu), un pour le sexe (M ou F) et l'autre pour le statut sanitaire (A ou S). Grâce à la fonction *data.frame()*, mettez vos vecteurs sous forme d'un tableau de données, dont chaque ligne représente un individu et chaque colonne une mesure prise sur un individu (dans notre cas, le sexe et le statut sanitaire). Enregistrez le tableau de données brutes dans une variable (par ex. donnees) et avec la fonction *summary()*, obtenez un résumé des données. La fonction *table()* permettra de créer une table de synthèse telle que reprise précédemment.

2. Dans R, générez les vecteurs abscisse et ordonnée qui vous permettront de tracer la courbe d'une distribution normale de moyenne 30kg et de déviation standard 3kg.