

LES TESTS D'HYPOTHESE

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 maîtriser les notions apprises au(x) TP(s) précédent(s).

Exercice 1

Si l'hypothèse généralement admise (hypothèse nulle ou H_0) est que les gens aiment autant avoir un chien qu'un chat (probabilité de préférence d'un chat = probabilité de préférence d'un chien = 0,5), on pourrait prédire la distribution des réponses à la question suivante, posées à 50 personnes : "préférez-vous posséder un chien plutôt qu'un chat ?". On peut calculer **dans excel** la probabilité que 20 personnes indiquent préférer les chiens dans Excel avec la fonction `LOI.BINOMIALE(nb succès; tirages; proba succès; cumulatif(vrai/faux))`, soit ici : `'=LOI.BINOMIALE.N(20;50;0,5;FAUX)'`. De même, la probabilité d'avoir 20 réponses ou moins en faveur des chiens s'écrit `'=LOI.BINOMIALE.N(20;50;0,5;VRAI)'`.

On pourrait aussi s'intéresser à une distribution alternative dans laquelle la probabilité de préférer le chien serait de $20/50 = 0,4$ et comparer cette distribution à celle de l'hypothèse nulle. Calculez ces deux distributions binomiales avec Excel, chacune dans une colonne, et faites-en une représentation graphique à l'aide de deux diagrammes en barres superposés. Pour cela, sélectionnez d'abord les deux plages de probabilités ainsi que les étiquettes éventuelles ' H_0 ' et ' H_1 ' puis cliquez sur Insertion -> Graphique... Choisissez le type de graphique et, dans les données sources, insérez la plage de valeurs de la variable binomiale (axe des abscisses).

N'oubliez pas que r , la variable binomiale, varie de 0 à n (soit 50 dans notre cas). Vérifiez que la somme des probabilités calculées pour chaque distribution vaut 100%. Etablissez ensuite la zone de rejet du test de l'hypothèse H_0 au seuil $\alpha = 5\%$ et calculez la puissance associée à ce test. Pour ce dernier calcul, déterminez d'abord la limite unilatérale r_{seuil} de r au seuil $\alpha = 5\%$ pour l'hypothèse nulle H_0 ($p = 0,5$). **Attention**, la valeur seuil (ici r_{seuil}) est choisie telle que la somme des probabilités de $r = 0$ jusqu'à $r = r_{\text{seuil}}$ vaut $\pm \alpha = 5\%$. Comme la distribution est discrète dans notre cas, on atteint rarement exactement 5%, il faut donc choisir r_{seuil} pour obtenir α aussi proche que possible de 0,05 sans jamais y être supérieure ($\alpha \leq 0,05$).

Si H_1 est vraie, la puissance du test de H_0 est la probabilité de rejeter - légitimement, donc - H_0 , c'est-à-dire la probabilité sous H_1 d'avoir, dans notre cas, une valeur de r plus petite ou égale à la valeur r_{seuil} calculée plus haut. Quelle est la puissance du test dans cette situation ? Et l'erreur β ?

Remarque : Dans vos calculs, pensez à utiliser des références relatives ; ainsi si vous essayez une probabilité alternative de 0,1, vous verrez les graphiques et les probabilités se modifier automatiquement. Comment évolue la puissance en fonction de la probabilité alternative ?

Appliquez ce même raisonnement **dans R** : déterminez la valeur seuil, l'erreur α puis l'erreur β et la puissance. Représentez graphiquement vos deux distributions sur un même graphique en utilisant `plot()` puis `lines()`. **Remarque** : vous pouvez jouer sur l'épaisseur des traits du graphique avec l'argument « `lwd=X` » et sur la continuité des traits avec « `lty=Y` » où X et Y sont des valeurs numériques entières.

Exercice 2

Dans R et dans Excel, calculez les valeurs des limites inférieure et supérieure de l'intervalle symétrique autour de la moyenne contenant 95% de la distribution théorique des poids des chiens de race Labrador. Les poids suivent une distribution normale de moyenne 30kg et de déviation standard 2kg. Si 95% des valeurs sont comprises (de manière symétrique autour de la moyenne) entre la limite inférieure et la limite supérieure trouvées, cela signifie que 2,5% des valeurs théoriques sont plus petites que la limite inférieure et 2,5% des valeurs théoriques sont plus grandes que la limite supérieure !

Dans R, échantillonnez 1000 chiens dans cette loi normale et trouvez les valeurs empiriques des limites inférieure et supérieure de l'intervalle 95%. Ces limites sont-elles similaires à celles obtenues de manière théorique ? Remarque : trie votre vecteur de valeurs avec la fonction `sort()`.

Exercices facultatifs pour le brain storming à domicile

1. A partir de l'exercice 2, faites une représentation graphique de la distribution empirique du poids des 1000 chiens échantillonnés et superposez-lui le graphique de la distribution théorique $N(30 ; 2^2)$.
2. Créez dans R un échantillon de 10000 valeurs issues d'une distribution continue uniforme comprise entre 10 et 60, arrondies à 3 décimales. Calculez-en la moyenne et la variance. Quelles étaient les valeurs attendues ? Ensuite, échantillonnez, hors de ces 10000 valeurs, 100 groupes de 10 données prises au hasard. Une même donnée peut être dans plusieurs groupes, il y aura donc remise. Pensez à la fonction `sample()`. Calculez la moyenne de chaque groupe et stockez ces 100 moyennes dans un vecteur. Calculez la moyenne et la variance de la distribution des moyennes des groupes de taille $n=10$. Comment devraient-elles évoluer ?

Faites une représentation graphique sur la même fenêtre des 2 distributions, celle des données brutes et celle des moyennes des groupes de taille $n=10$. Le graphique colle-t-il bien avec les résultats que vous avez obtenus pour les moyennes et variances ? Quelle est la forme de la distribution des moyennes des groupes.