Statistique et Informatique — UFR STAPS
Licence 3 — Université de Paris X - Nanterre

TD4
Loi Normale et intervalles de confiance:

1 Du pourcentage à l'intervalle

Dans un lycée super sympa, les élèves ont une moyenne générale de 8,32 écart type 2,67. Le proviseur décide pour améliorer les résultats, de récompenser les bons et d'aider les moins bon. Mais avant, il veut faire une petite étude statistique pour voir le coût d'une telle opération.

Rappel : Pour trouver un intervalle a partir d'un pourcentage p, il faut :
– trouver le z, soit sur la table, soit en utilisant la formule `z <- qnorm(p,lower.tail=FALSE)`
– L'intervalle contenant le pourcentage p est alors au choix soit $[m + s \times z; +\infty[$, soit $]-\infty; m - s \times z[$.

1. Quel intervalle contient les 2.5% meilleurs ?
2. Quel intervalle contient les 2.5% moins bons ?
3. Pour recevoir un livre, il faut être dans le top 20% du lycée. Donnez l'intervalle contenant le top 20% des élèves.
4. Parallèlement, on propose des cours de soutien aux étudiants en difficulté. Quel intervalle contient les 30% moins bonnes notes ?

Les cas suivant sont un peu plus compliqués et pour les résoudres, il faut faire un schéma sur une feuille à part :

– Dessinez approximativement une loi normale
– Sur l'axe des x, placez la moyenne de la loi normale.
– Grisez la surface que l'on cherche. Point important qui doit apparaitre sur le graphique, cette surface contient-elle la moyenne ?
– Tracez sous la surface un trait représentant l'intervalle.
– Décomposez votre intervalle en une somme ou une soutraction d'intervalles pour lesquels il est possible de trouver z. Rappels : seul les intervalle de la forme $]-\infty; z]$ si z est plus petit que la moyenne ou $[z; +\infty[$ si z est plus grand que la moyenne sont calculables.

5. L'intervalle centré contenant 95% d'une population (et excluant les 2.5% plus haut et plus bas indicidus) s'appelle *intervalle de confiance au risque 5%*. Sans faire de calcul supplémentaire, quel est intervalle de confiance des lycéens au risque 5% ?
6. Les cours de soutien sont vraiment très profitables. On généralise le principe et on le propose à 80% des élèves. Sans faire de calcul supplemantaire (mais avec l'aide d'un bon dessin), quel intervalle contient les 80% moins bonnes notes ?

2 De l'intervalle au pourcentage

Dans tout ce qui précède, nous partions d'un pourcentage p et nous trouvions l'intervalle grâce à la formule $x = m + s.z$. Il est possible de faire le chemin inverse et trouver un pourcentage correspondant à un intervalle de la population. Pour cela, il suffit d'inverser la formule : $z = \frac{x-m}{s}$ (ou $z = \frac{m-x}{s}$ quand x est plus petit que m).

Rappel : Pour trouver un pourcentage à partir d'un z, vous pouvez utiliser soit sur la table, soit la formule `p <- pnorm(z,lower.tail=FALSE)`

Dans ce même lycée, la moyenne des terminales au bac est de 9,24 écart type 2,3.

7. Quelle est la probabilité d'avoir 12 ou plus (une mention) ?
8. Quelle est la probabilité d'avoir 16 ou plus (au moins mention Trés Bien) ?
9. Quelle est la probabilité d'avoir moins de 8 ?

La encore, pour les intervalles un peu plus compliqués, il faut faire un schéma. Pour cela :
– Dessinez approximativement une loi normale
– Sur l'axe des x, placez la moyenne de la loi normale.
– Sur l'axe des x, placez les bornes de l'intervalle. Si une des bornes est infinie, placez la sur le bord gauche ou droit de l'axe des x.
– Tracez un trait entre les deux bornes que vous venez de placer pour représenter l'intervalle.
– Décomposez votre intervalle en une somme ou une soutraction d'intervalles pour lesquels il est possible de trouver p. Rappels : les seuls intervalles pour lesquels il est possible de calculer p sont de la forme $]-\infty; z]$ si z est plus petit que la moyenne ou $[z; +\infty[$ si z est plus grand que la moyenne.

10. Quelle est la probabilité de ne pas avoir le bac à la premiere session ?
11. Sans faire de calculs supplémentaires, quelle est la probabilité d'être convoqué au rattrapage ?
12. Si Stéphanie à 16 ou plus, elle intégre la prestigieuse prépa Saint Louis. Si elle a entre 12 et 16, elle va à l'université. Quelle est la probabilité que Stéphanie aille à l'université ?
13. Quel est la probabilité d'avoir entre 9 et 11 ?

3 Vérification des propriétés de la loi normale

En CM, on vous a raconté qu'on pouvait construire des intervalles et deviner le nombre de personnes qui s'y trouvent simplement en sachant qu'une variable suit une loi normale. Est-ce vrai ?

3.1 Data.frame

14. Un fichier nommé *L3StatNotes2006.csv* se trouve sur le site Internet. Enregistrez-le dans le répertoire que vous avez créé dans *Mes Documents*
15. Importez ce fichier sous **R** et stockez le dans la variable `donnees` (n'oubliez pas : sauvegarde du fichier au bon endroit, changer le repertoire courant, `read.csv2()`,
16. Après avoir ouvert un fichier contenant un data.frame et l'avoir stocké, la première chose à faire est de vérifier son contenu. Pour cela, on utilise la commande `summary()`. Tapez `summary(donnees)`. Qu'obtenez-vous ?
17. Déterminez (à la main, sans **R**) la liste des variables et leur type (à noter dans votre fichier Word).

3.2 Analyse univarié sommaire

18. Calculez la moyenne et l'écart type (grâce aux fonctions `mean()` et `sd()`) tout en ignorant les manquantes avec `na.omit()` de `[MiniQCM]`, `[Exam]` et `[NoteFinale]`
19. Tracez les histogrammes de ces variables. Suivent-elles une loi normale ?

3.3 La variable `[MiniQCM]`

Dans tout ce qui suit, on suppose que la variable `[MiniQCM]` suit une loi normale. On vous rappelle que vous connaissez la moyenne et l'écart type de cette loi puisque vous les avez calculés juste au dessus pendant l'analyse univariée.

Note : **R** peut vous aider à faire les calculs (comme le ferait une calculatrice), mais il n'existe pas de formule vous donnant les pourcentages automatiquement. Désolé !

20. Quel est le nombre total d'étudiant de licence ? Combien d'étudiant sont manquants pour `[MiniQCM]` ? Au final, combien ont une note ? Stockez ce nombre dans `presentMiniQCM`.
21. Recalculez la moyenne et l'écart type de `[MiniQCM]` et stockez-les dans les variables `moyMiniQCM` et `ecartMiniQCM`.

22. Quel pourcentage d'étudiant doit théoriquement être dans l'intervalle [12; 20] ? Pour z, faites un calcul exact en utilisant `moyMiniQCM` et `ecartMiniQCM`, puis consultez la table de la loi normale sur le site internet.
23. **R** est capable de calculer le pourcentage associé a un z donnée grâce à la fonction `pnorm(z,lower.tail=FALSE)`. L'avantage de cette méthode est qu'elle est plus précise. Retrouvez le pourcentage précédent en utilisant **R** et stockez le résultat dans `pMiniQCM`.
24. En vous basans sur le nombre d'étudiant ayant une note et sur le pourcentage d'étudiant ayant théoriquement entre 12 et 20, combien d'étudiant sont théoriquement dans l'intervalle [12; 20] ?
25. **R** permet de compter automatiquement les individus qui ont plus de 12. Pour cela, il faut trouver tous les gens qui ont plus de 12, puis enlever les valeurs manquantes et enfin faire la somme de ceux qui restent. L'instruction permettant de faire ça est `sum(na.omit(donnees$MiniQCM >= 12))`. (vous pouvez faire un copier-coller depuis ce document jusqu'à **R**). Note : Cette instruction étant hors programme, je ne la détaille pas. Néanmoins, si vous souhaitez en comprendre le fonctionnement, n'hésitez pas à me demander.

Vous avez trouvez deux chiffres. Le premier, celui de la question 24, est une prédiction à l'aide de l'intervalle de confiance. Le deuxième (question 25), est issu de la réalité.

26. Que pensez-vous de la prédiction faite grâce à l'intervalle de confiance ?

Comme toujours avec **R** la première fois est longue, les autres sont rapides :

27. Quel pourcentage d'étudiant doit théoriquement être dans l'intervalle [14; 20] ?
28. Vérifiez en comptant (grâce à `sum(na.omit(donnees$NoteFinale >= 14))`).

3.4 La variable [Examen]

On suppose que [Examen] suit une loi normale.

29. Quel pourcentage d'étudiant est dans l'intervalle [12; 20]. Stockez la réponse dans `pExamen`.
30. Combien d'étudiant ont une note à l'examen ? Stockez la réponse dans `presentExamen`.
31. Combien d'étudiant doivent théoriquement avoir plus de 12 à l'examen ?
32. Vérifiez en comptant (grâce à `sum(na.omit(donnees$Examen >= 12))`).
33. La prédiction faite grâce à l'intervalle de confiance est-elle bonne ? Comment l'expliquer ?

3.5 Conclusion

34. Dans quel cas, sous quelles conditions peut-on faire de bonnes prédictions ?