Cours1_Bootstrap

January 19, 2024

1 Bootstrap :

des intervalles de confiance flexibles et non-paramétriques En statistique *inférentielle*, on extrapole des observations faites sur un échantillon à l'ensemble de la population dont l'échantillon est issu. Ce faisant, on prend un risque de se tromper et donc on doit aussi quantifier notre incertitude quant à notre extrapolation.

En statistique, l'incertitude, c'est de la variance : Si l'on répétait une multitude de fois i) la même acquisition de données (= échantillonnage de la population) et ii) analyse de l'échantillon, quelle serait la variabilité de la valeur observée à chaque analyse ?

Lorsque l'analyse statistique de l'échantillon consiste simplement à calculer la moyenne de la variable d'intérêt, et que l'échantillon contient plus de 30 observations i.i.d., le théorème central limite s'applique. Ce théorème nous enseigne que si l'on répétait une multitude de fois l'échantillonnage, l'ensemble des moyennes calculées sur chaque échantillon suivrait une loi normale centrée sur la moyenne de la population, parfois appelée 'moyenne vraie', parce que c'est la moyenne que l'on cherche à estimer avec notre échantillon. La variance de ces moyennes représente l'incertitude que l'on a lorsque l'on calcule une seule moyenne basée sur un échantillon. Le théorème central limite nous enseigne que cette variance est de $\sigma^2_\mu = \frac{\widehat{\sigma^2_{pop}}}{n}$, où $\widehat{\sigma^2_{pop}}$ correspond à la variance de la population estimée ($\widehat{\ }$) à l'aide de l'échantillon, et n, la taille de l'échantillon. Connaitre ces propriétés de cette loi normale permet de calculer, à l'aide d'un seul échantillon, un intervalle de confiance à seuil de risque souhaité, typiquement 5%.

Lorsque l'on souhaite calculer un intervalle de confiance pour une autre statistique que la moyenne, **ou/et que** le nombre d'observations est inférieur à 30, on ne peut pas utiliser le théorème central limite pour calculer un intervalle de confiance pour notre statistique. À la place, on peut utiliser les bootstraps.

Dans les grandes lignes, la logique sous-jacente aux bootstraps est la même que celle du théorème central limite : * On souhaite connaitre quelle est la distribution qu'aurait la statistique qui nous intéresse (par exemple la médiane) si l'on répétait l'acquisition de donnée et l'analyse une multitude de fois. * L'incertitude doit être d'autant plus grande que le nombre de données n est faible. * Dans le cadre du théorème central limite, on a utilisé ces données, notre échantillon, pour approximer les variances de la population et de la statistique (σ^2_{pop} et σ^2_μ). Dans le cadre des bootstraps, on utilise notre échantillon, pour approximer la distribution de la population et celle de notre statistique d'intérêt :

1. **Approximer la distribution de la population** Si dans notre échantillon, on a notamment 3 fois la valeur 5 et 1 fois la valeur 11, ceci signifie que la *meilleure* ***estimation*** *que l'on a* de la fréquence relative des valeurs 5 et 11 dans la population est qu'il y a 3 fois plus de 5

que de 11. Si l'on souhaite ***simuler*** un autre échantillonnage dans la population, sans avoir les vraies fréquences relatives de ces valeurs dans la population, *le mieux que l'on puisse faire* est d'utiliser les fréquences relatives observées dans notre échantillon. Il s'avère que, plutôt que de calculer les fréquences relatives de chacune des valeurs observées dans l'échantillon pour ensuite ***simuler*** un ré-échantillonnage dans la population, il est mathématiquement strictement identique et beaucoup plus simple d'effectuer cette simulation en échantillonnant avec remise dans notre propre échantillon réellement observé.

2. **Approximer la distribution de notre statistique** Ainsi, en échantillonnant avec remise dans nos données, on peut simuler autant d'échantillonnages dans la population qu'on le souhaite, et de la taille qu'on le souhaite. Mais étant donné qu'on vise *in fine* à estimer notre incertitude dans notre statistique, elle-même estimée grâce à notre échantillon réel de taille n, on réalise des ré-échantillonnages de taille n. Sur chacune de ces *simulations d'échantillonnage dans la population*, on recalcule notre statistique d'intérêt. En première approche, on peut calculer l'intervalle de confiance de notre statistique via les quantiles des statistiques simulées par ré-échantillonnage de nos données (méthode dite des 'percentiles'). Par exemple pour un intervalle de confiance à 95%, on va utiliser comme borne inférieure et supérieure, les quantiles $Q_{0.025}$ et $Q_{0.975}$.

En résumé, on procède comme suit : > 1. *Simulez* de très nombreux (B) échantillonnages dans la *population*, en échantillonnant avec remise dans l'échantillon réel, observé : nos données. On réalise typiquement 2000 à 10000 simulations ; plus on souhaite un niveau de confiance de notre intervalle fort, proche de 100%, plus le nombre de simulations doit être grand. Il faut typiquement viser $B = \frac{100}{1-\text{niveau de confiance}}$ simulations, soit pour un intervalle de confiance à 95% : $B = \frac{100}{1-\frac{95}{100}} = 2000$ simulations, et à tout prix, il faut éviter d'être en deçà de $B = \frac{30}{1-\text{niveau de confiance}}$ simulations. > 2. Sur chacune de ces B simulations, calculez la statistique d'intérêt. L'ensemble de ces B statistiques simulées constitue notre *estimation* de la distribution qu'auraient nos statistiques si nous répétions réellement l'échantillonnage une multitude de fois. > 3. Utilisez cette distribution estimée de notre statistique pour estimer son intervalle de confiance. pour ce faire, la méthode la plus simple repose sur l'utilisation des percentiles, mais d'autres approches existent. Pour plus de détails, voir : liens 1

Limites : Les bootstrap sont globalement très robustes et versatiles. Notamment, lorsque les données ne sont pas i.i.d., il est souvent possible d'adapter la procédure d'échantillonnage avec remise pour prendre en compte la structure de dépendance dans les données (lien). Cependant, les hypothèses sont que * L'échantillon est représentatif de la population dont il est issu - en d'autres termes, il contient suffisamment d'informations et sa sélection n'a pas été biaisée. Ainsi, ***contrairement à ce qui est fréquemment dit***, les bootstraps ne sont pas forcément adaptés pour des petits échantillons : **cette approche ne fonctionne pas bien si des événements rares sont absents de l'échantillon étudié, et affectent fortement la statistique utilisée**. * La théorie asymptotique doit être applicable : la statistique utilisée sur l'échantillon doit être un bon estimateur de la statistique appliquée la population. Cette applicabilité de la théorie asymptotique dépend à la fois de l'estimateur et de la distribution sur laquelle elle est appliquée. Ce problème peut être détecté en s'intéressant au ***biais*** (voire ci-après). * La population est infinie ou suffisamment grande pour que l'effet du prélèvement d'un échantillon soit négligeable ($\Rightarrow$ donc on peut simuler d'autres échantillonnages dans la population en échantillonnant avec remise dans la population).

1.1 Last but not least, le problème de bias :

La distribution bootstrap et l'échantillon peuvent être *systématiquement* en désaccord : $E(S_{boot}) \neq S_{obs}$ Ce qui traduit que le même problème existe pour notre échantillon par rapport à la population : $E(S_{obs}) \neq S_{pop}$

Illustration concrète :

[ ]:
```
## la statistique de la médiane pour la popluation (un TRÈS grand échantillon ;␣
 ↪n=1e6)
median(rexp(1e6, 1))
# 0.3474317

## l'espérance de la statistique de la médiane pour des petits échantillons␣
 ↪(n=10)
mean(
  sapply(1:1e4, function(r)
    median(rexp(10, 1))
  )
  )
# 0.7477992
```

Dans ce cas, un biais peut se produire. Si la distribution bootstrap d'un estimateur est symétrique, des intervalles de confiance percentiles sont souvent utilisés ; ces intervalles sont particulièrement appropriés pour les estimateurs sans biais. Dans le cas contraire, si la distribution de bootstraps n'est pas symétrique, les intervalles de confiance basés sur les percentiles sont souvent imprécis, car eux-mêmes affectés par le biais. Il est alors préférable d'utiliser le 'bias-corrected and accelerated (BCa)' bootstrap.

/!\ : Ces méthodes BCa et dérivées ne corrigent pas pour le biais entre la statistique observée (S_{obs}) et les statistiques obtenues sur les bootstraps (S_{boot}). Elles utilisent ce biais ($S_{obs} - S_{boot}$) pour estimer le biais entre S_{obs} et S_{pop} et corriger ce dernier afin d'estimer S_{pop} et son interval de confiance. Pour plus de détails, voir ce lien.

Ce problème du biais permet aussi de pointer une autre utilisation possible des bootstrap : Vérifier si la statistique S_{obs} est un bon estimateur de S_{pop}.

La taille minimale de l'échantillon pour appliquer des bootstraps dépend de i) la statistique utilisée, ii) de la distribution de la population (présence de valeurs extrêmes rares impactant la statistique ou non), iii) le seuil désiré pour l'intervalle de confiance.

De ce fait, il est compliqué de donner un effectif minimal requis, mais les valeurs qui circulent varient de 8 à 10, à 50 à 60 … Et elles sont loin d'être toujours respectées.

Gardez à l'esprit que les bootraps sont très puissants, mais qu'ils ne sont pas magiques. Si les données sont faibles ou de mauvaises qualités, ou que la statistique utilisée est inappropriée, vos résultats ne seront jamais mirobolants (sauf si le prestidigitateur Didier Raoult est dans votre équipe !).

Quelques ressources pour aller plus loin : Il est aussi possible d'utiliser les boostraps pour calculer des *p*.values : lien

Deux autres explications intuitives des bootstraps : 1 et 2

2 tutoriels pour faire des bootstrap en R : 1 et 2

Un super cours sur les bootstrap

Recent Developments in Bootstrap Methodology (2003)

Détails sur les bootstrap 'bias-corrected and accelerated (BCa)' : 1 et 2

Des biais… Is it true that the percentile bootstrap should never be used ?

Why not always use bootstrap ?