Le rôle de la statistique dans nos vies: démonstration de multiples applications au quotidien

L'importance de la cible

Christian Léger

Université de Montréal

29 novembre 2013

Table des matières

Université de Montréal

Année mondiale de la statistique

Université de Montréal

Entrevue à RDI : *Joint Statistical Meetings*

Université de Montréal

- Le plus grand congrès mondial de statistique a attiré 6000 statisticiens à Montréal en août dernier dans le cadre de l'année mondiale de la statistique.
- Question
- De quoi peuvent-ils bien parler (de combien de manières différentes peut-on calculer une moyenne ?)
- « Ne vous surprenez pas cette semaine de voir plein de gens avec des calculatrices dans leurs mains, plus de 6000 statisticiens sont en ville »...

Qu'est-ce la statistique ?

Université de Montréal

- La statistique : ce n'est pas d'abord et avant tout la manière de résumer les résultats sportifs (malgré *Moneyball* et l'intérêt porté par une équipe de statisticiens à Harvard)...
- et c'est beaucoup plus qu'un cours ennuyant que vous avez suivi au CEGEP ou à l'Université...
- La statistique est l'art et la science d'extraire l'information de données.

Importance de la statistique

Université de Montréal

- La chimie, la physique, ou le génie civil, pour ne nommer que ces disciplines, ont clairement plein d'applications dans la vie de tous les jours.
- Mais certainement pas la statistique, n'est-ce pas ?
  - De la prise de vos médicaments le matin
  - au taux de chômage que vous entendez à la radio le matin,
  - aux légumes que vous mangerez ce midi,
  - à la demande de prêt personnel ou de carte de crédit que vous ferez cet après-midi,
  - au choix des sites que Google vous suggérera suite à une requête en soirée
  - aux prévisions météo que vous regarderez en fin de soirée,
  - la statistique y aura joué un rôle important !

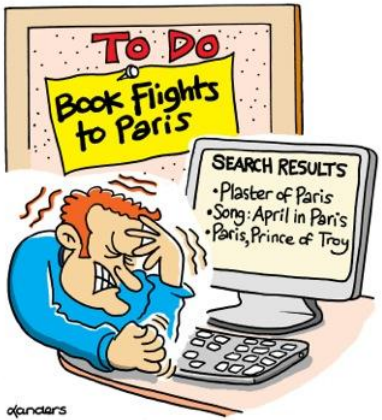
Exemples de nouveaux défis en statistique

Université de Montréal

- Puissance de calcul.
- *Big Data*.
- Nouvelles technologies : multiplicité des tests avec les expériences de micropuces.
- Nouvelles technologies : données de déplacement d'animaux marins.
- Modélisation pour données longitudinales.
- Données textuelles.

Recherche sur Google

Université de Montréal



Without Statistics we would have to search the internet one site at a time.

Coupons de Target

Université de Montréal

Différentes étapes où le statisticien intervient

Université de Montréal

- La planification d'une expérience ou d'un plan de sondage.
  - Comment planifier une expérience pour déterminer si un nouveau médicament est supérieur au traitement actuel ?
  - Quel plan de sondage doit-on utiliser pour mesurer le taux de chômage au Canada ?
  - Comment poser des questions qui vont mener à des réponses objectives ?
- La collecte des données.
  - Comment maximiser le nombre de personnes qui vont répondre à l'enquête afin de minimiser la non-réponse ?
  - Comment s'assurer du bon déroulement d'une expérience afin que les résultats puissent mener à des conclusions valables ?
  - Détecter des erreurs lors de l'entrée de données ?

Différentes étapes où le statisticien intervient

Université de Montréal

- L'analyse des données.
  - C'est le gros de notre travail.
  - Les mêmes outils peuvent être utilisés dans des applications complètement différentes : quelle est la probabilité qu'une personne ne rembourse pas son prêt ou quelle est la probabilité qu'une personne décède d'une certaine forme de cancer.
  - La méthodologie évolue constamment !
- L'interprétation des résultats.
  - Quelles conclusions doit-on tirer des résultats ?
  - Ça dépend entre autres de la planification faite au début.

Introduction : conclusion

Université de Montréal

Message : La statistique est une science très vivante, en constante évolution. Elle tire une grande partie de son innovation de son interaction avec les autres disciplines. Elle est omniprésente.

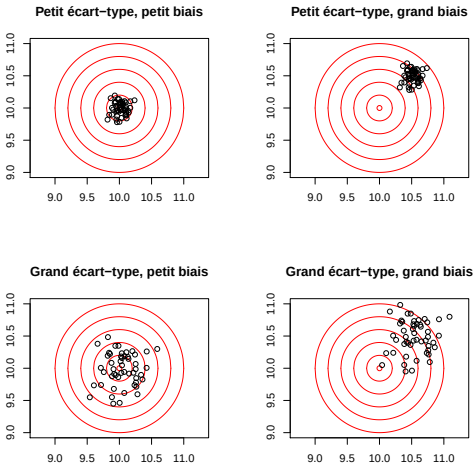
Table des matières

Biais et variance d'un estimateur

Université de Montréal

- Temps moyen pour se rendre au travail T et au salaire moyen S.
- Un échantillon de personnes et pour chaque personne on mesure le temps pris pour se rendre au travail et son salaire. On dénote par $\hat{T}$ et $\hat{S}$ le temps moyen et le salaire moyen calculés sur l'échantillon.
- $\hat{T}$ et $\hat{S}$ sont des estimateurs de T et S, des paramètres de la distribution.
- Deux concepts pour évaluer un estimateur sont le biais et l'écart-type.
- Biais : Si la moyenne de la distribution de l'estimateur (c-à-d de toutes les valeurs que pourrait prendre l'estimateur) est le paramètre qu'il estime, on dit de l'estimateur qu'il est sans biais. Sinon, la différence entre cette moyenne et le paramètre est le biais.
- L'écart-type d'un estimateur est une mesure de sa dispersion.

Biais et écart-type d'un estimateur



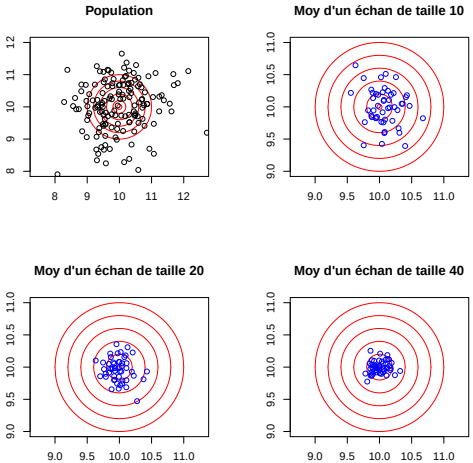
Test pour l'égalité de deux moyennes

Université de Montréal

- On veut tester l'hypothèse d'égalité de la moyenne de deux groupes.
- Modèle : $X_1, \ldots, X_m$ sont les variables aléatoires indépendantes et identiquement distribuées du premier groupe de taille m et $Y_1, \ldots, Y_n$ sont les variables aléatoires indépendantes et identiquement distribuées du second groupe.
- On postule que la moyenne des X_i est μ_X et que celle des Y_i est μ_Y alors que l'écart-type commun des X et des Y est σ.
- *Sous les hypothèses du modèle*, on peut montrer que l'écart-type de la moyenne de l'échantillon, $\sigma_{\bar{X}}$, est l'écart-type de la population (distribution) divisée par la racine carrée de la taille de l'échantillon, $\sigma/\sqrt{m}$.
- Ainsi, on comprend bien le comportement de la dispersion de la moyenne de l'échantillon.

Dispersion de la population et de la moyenne d'un échantillon

Université de Montréal



Test pour l'égalité de deux moyennes

Université de Montréal

- Sous les mêmes hypothèses citées précédemment, la moyenne de l'échantillon $\bar{X} = (1/m)(X_1 + X_2 + \ldots X_m)$ est sans biais pour la moyenne de la population μ_X, c-à-d que la moyenne de la distribution des $\bar{X}$ (si on recommençait l'expérience plusieurs fois) est la moyenne de la population.
- Le test pour l'égalité des deux moyennes μ_X et μ_Y rejette lorsque la statistique T est très petite (négative) ou très grande (positive) où

$$T = \frac{\bar{X} - \bar{Y}}{s\sqrt{1/m + 1/n}}$$

et s est l'écart-type combiné des deux échantillons.

- Sous les hypothèses du modèle et si les données sont distribuées sous une distribution normale, on peut montrer que si les moyennes sont identiques, la distribution de la statistique T est une distribution t de Student à $m + n - 2$ degrés de liberté.

Peut-on toujours utiliser la distribution t ?

Université de Montréal

- Si j'ai deux ensembles de données et que je veux tester l'hypothèse de l'égalité des deux moyennes, est-ce que je peux toujours utiliser la statistique T et la distribution t à $m + n - 2$ degrés de liberté ?
- Si les données ne sont pas normales, mais que la taille des échantillons est grande, ça fonctionnera quand même.
- Si la dispersion dans les deux groupes est différente, on peut modifier le dénominateur et ajuster les degrés de liberté.
- Mais si les données ne sont pas indépendantes, alors ça ne tient plus et ça prend des ajustements plus importants qui vont dépendre d'un modèle (séries chronologiques, effet de grappe, etc.), mais on sait quoi faire.

Peut-on toujours utiliser la distribution t ?

Université de Montréal

- Si je pose mes questions aux 10 premiers travailleurs à se présenter à l'usine le matin, est-ce représentatif de tous les travailleurs de l'usine ?
- Échantillonnage par grappes : On choisit d'abord un certain nombre de quartiers au hasard, puis à l'intérieur de chacun d'eux, on choisit un échantillon de personnes dans le quartier.
- Westmount vs Hochelaga-Maisonneuve vs partout.
- Traiter plusieurs observations sur un même patient comme étant des observations indépendantes est une sérieuse erreur commune : plusieurs observations sur différentes parties du corps comme en arthrite rhumatoïde. Dans les années 80 une étude de 196 essais cliniques de recherche en arthrite rhumatoïde a démontré que 63% d'entre elles avaient mal analysé les observations multiples sur les mêmes patients !

Dispersion : conclusion

Université de Montréal

Message : Les statisticiens comprennent plutôt bien la dispersion d'un estimateur. Grâce à un modèle, on peut l'estimer à partir des données disponibles. Et on peut souvent vérifier l'adéquation du modèle postulé.

Table des matières

Recensement

Université de Montréal

- Le recensement au Canada a lieu aux 5 ans (en xxx1 et xxx6).
- Jusqu'en 1971, tous les ménages canadiens remplissaient un questionnaire contenant un grand nombre de questions.
- En 1971, on a décidé que seulement un cinquième des ménages rempliraient le formulaire long, les autres remplissant le formulaire court d'environ 10 questions.
- Le recensement (formulaires long et court) est **obligatoire** pour tous.

Enquête nationale des ménages

Université de Montréal

- En 2010, le gouvernement Harper a décidé de remplacer le formulaire long (**obligatoire**) du recensement de 2011 par l'Enquête nationale des ménages (ENM) qui était **volontaire**, sous prétexte que le formulaire long était une intrusion de la vie privée.
- La communauté scientifique et la société civile se sont fortement prononcés contre, mais sans succès.
- Ils prédisaient que le taux de réponse serait grandement diminué, ce qui créerait un biais.
- Pour pallier au problème, le gouvernement a décidé que Statistique Canada distribuerait l'ENM à 30% des ménages plutôt qu'à 20%.

Problème de la non-réponse

Université de Montréal

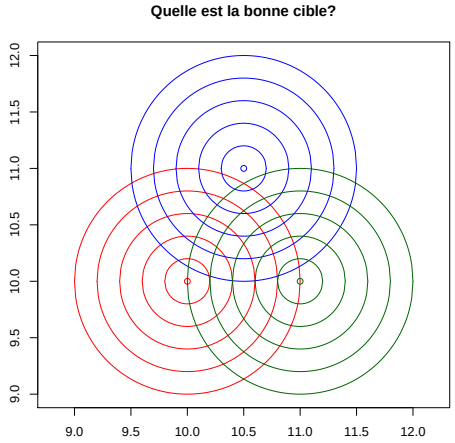
- Lorsqu'il y a de la non-réponse, la taille de l'échantillon diminue, ce qui diminue la précision des estimateurs (l'écart-type augmente).
- Dans bien des cas, un biais peut apparaître de telle sorte qu'en moyenne, l'estimateur n'estime plus la bonne quantité.
- Si on a un modèle pour la non-réponse, on peut tenter de modifier l'estimateur pour tenter d'éliminer le biais...
- ...toutefois il est plus facile de gérer l'écart-type que le biais.

Simulation

Université de Montréal

- Afin de mieux saisir les difficultés associées à la non-réponse, nous allons procéder à une simulation de lancer de dards.
- La position des dards représente la valeur de deux variables (ses coordonnées en x et y).
- Si tous les dards étaient sur la cible, on définirait le centre de la cible comme étant la moyenne de la position de tous les dards ; c'est donc le paramètre qu'on cherche à estimer (cette position est inconnue).

On ne connait pas la cible



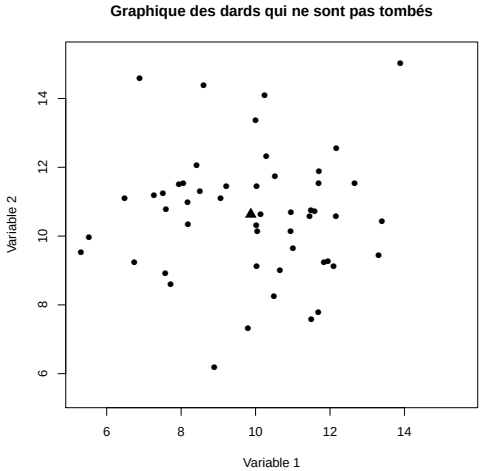
Simulation

Université de Montréal

- On prend un échantillon de dards.
- Parce qu'ils n'avaient pas suffisamment de vélocité ou pour d'autres raisons, des dards tombent par terre : c'est comme la non-réponse, nous n'avons pas la position de ces dards.
- Comment estimer le centre de la cible à ce moment-là ?
- Regardons un échantillon de 100 dards.
- Nous n'observons pas tous les dards : il y a de la non-réponse. Le triangle en noir représente la moyenne des dards restés sur le mur (cercles pleins) alors que le triangle bleu est la moyenne de tous les dards, incluant ceux qui sont tombés (les cercles vides dans le deuxième graphique). Le triangle noir est notre **estimation** du centre de la cible, qui est le **paramètre**.

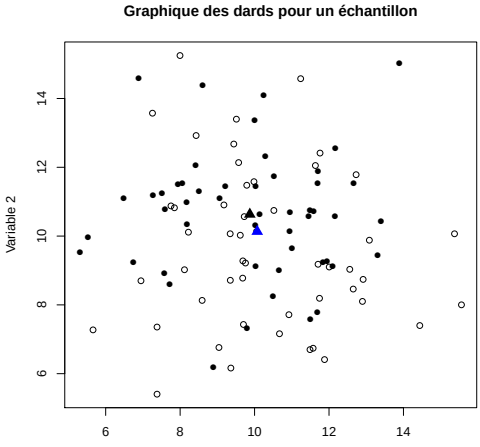
Graphique des dards qu'on peut observer

Université de Montréal



Graphique des dards tombés et ceux sur le mur

Université de Montréal



Simulation de plusieurs échantillons

Université de Montréal

Nous venons de présenter un échantillon de 100 dards de notre modèle avec un certain nombre d'entre eux qui sont tombés. Ceci nous mène à *une* moyenne de tous les dards (qu'on ne peut pas observer à cause de la non-réponse) et *une* moyenne des dards qui ne sont pas tombés (qu'on peut observer). Ça ne nous donne pas vraiment d'information sur ces estimateurs. Il nous faudrait *plusieurs* valeurs des estimateurs pour mieux comprendre.

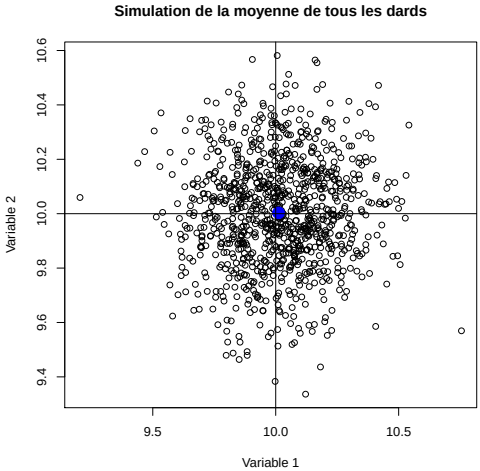
Simulation de plusieurs échantillons

Université de Montréal

On simule 1000 échantillons et pour *chaque échantillon* on calcule la moyenne de tous les dards (qu'on ne pourrait pas calculer en réalité) et celle des dards qui ne sont pas tombés.

Dans le graphique suivant, chaque point représente la moyenne de tous les dards d'un échantillon (le triangle bleu dans les graphiques précédents). La moyenne de tous ces points par un gros cercle bleu. L'intersection des deux droites est le centre de la cible.

La moyenne de la moyenne de tous les dards



La moyenne de tous les dards est sans biais

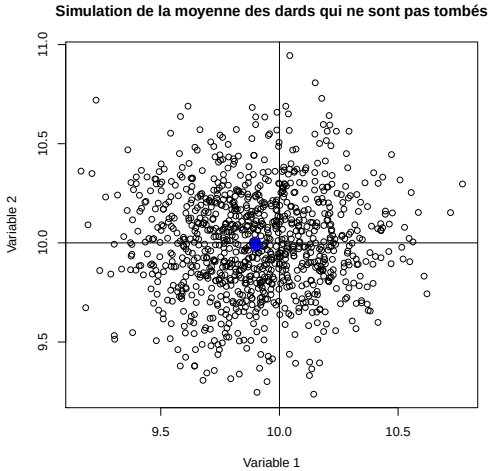
Université de Montréal

Le cercle est *essentiellement* sur la cible. Ceci n'est pas surprenant puisque la moyenne d'un échantillon i.i.d. est sans biais pour la moyenne d'une distribution.

Mais on ne peut pas calculer cet estimateur puisque nous n'avons pas le résultat des dards qui sont tombés.

Qu'en est-il de la moyenne des dards qui ne sont pas tombés ?

La moyenne de la moyenne des répondants

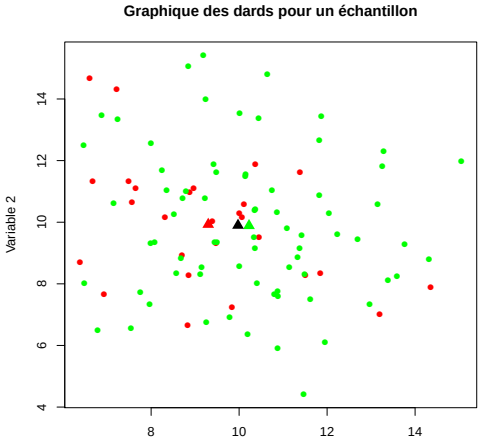


La moyenne des dards observés est **biaisée**

Université de Montréal

- Le cercle est bien à côté de la cible : Si on prend la moyenne des répondants (les dards qui ne sont pas tombés), **en moyenne** l'estimateur sera à gauche de la cible. L'estimateur est donc biaisé en x (mais pas en y).
- Que se passe-t-il ?
- Nous avons de l'information supplémentaire sur les dards, à savoir s'ils ont été lancés par un droitier ou un gaucher.
- Les dards des droitiers sont en vert et ceux des gauchers en rouge.

Droitiers vs gauchers



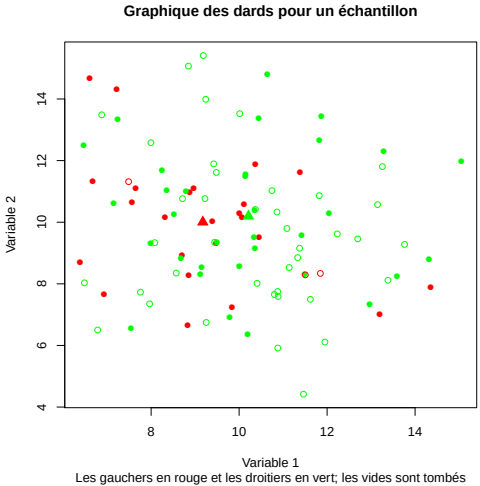
Les gauchers en rouge et les droitiers en vert; les triangles sont les moyennes

Droitiers vs gauchers

Université de Montréal

- La moyenne des droitiers (triangle vert) est à droite de celle des gauchers (triangle rouge) ; le triangle noir est la moyenne de tous les dards.
- **Si** les droitiers avaient tendance à lancer plus à droite et les gauchers plus à gauche, les moyennes des droitiers et des gauchers seraient compatibles avec celles observées.
- Mais des dards sont tombés. . .

Droitiers vs gauchers



Droitiers vs gauchers

Université de Montréal

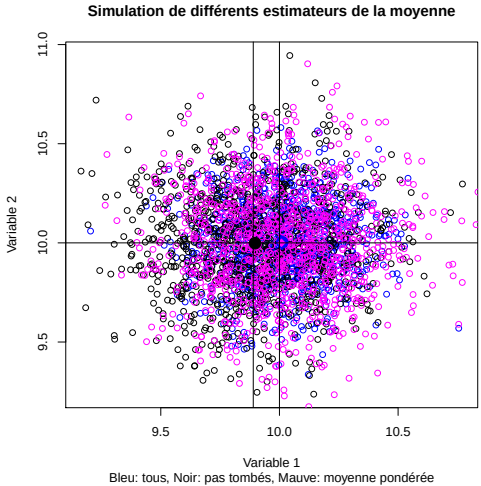
- Que remarque-t-on ?
- Il y a une plus grande proportion de dards lancés par des droitiers qui sont tombés que chez les gauchers.
- La probabilité de répondre n'est pas la même pour tous : elle est plus grande chez les gauchers que chez les droitiers.
- (Si tous les gauchers ont la même probabilité de répondre et proviennent de la même distribution) la moyenne des dards des gauchers qui ne sont pas tombés est sans biais pour la moyenne de la distribution des gauchers. Et de façon similaire pour les droitiers.

Un estimateur sans biais

Université de Montréal

- Si on connaît la proportion de droitiers et de gauchers dans la population (comme on pourrait la connaître via un recensement), on pourrait définir un nouvel estimateur $\hat{\theta} = p_g \bar{X}_g + p_d \bar{X}_d$ où p_g et p_d sont les proportions de gauchers et de droitiers *dans la population* et $\bar{X}_g$ et $\bar{X}_d$ sont les moyennes des répondants gauchers et droitiers *dans l'échantillon des répondants*.
- Dans le prochain graphique, on représente la valeur de 1000 réalisations de trois estimateurs : la moyenne de toutes les observations (si tout le monde répondait) en bleu, la moyenne des répondants en noir et la moyenne pondérée des droitiers et gauchers en mauve.
- On remarque que la moyenne de toutes les observations et la moyenne pondérée sont sans biais alors que la moyenne des répondants est biaisée.

L'estimateur pondéré $\hat{\theta}$ est sans biais

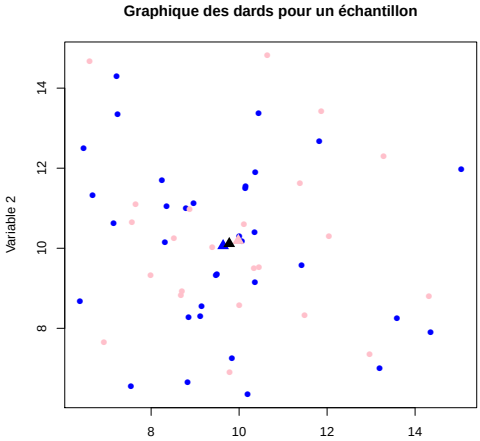


Homme vs femme

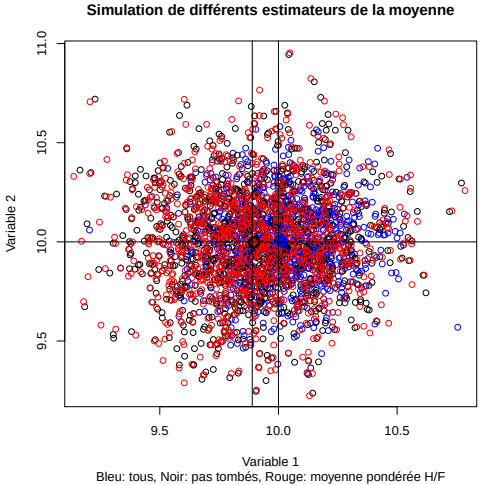
Université de Montréal

- Supposons qu'on connaisse aussi le sexe de la personne qui lance le dard : peut-être les hommes lancent-ils les dards plus haut et les femmes plus bas, ce qui pourrait mener à un biais si les dards des hommes et des femmes tombaient avec des probabilités différentes.
- Dans ce modèle, nous avons une probabilité de 0,6 que le dard soit lancé par un homme.
- Un autre estimateur devient $0{,}6\bar{X}_H + 0{,}4\bar{X}_F$ où $\bar{X}_H$ et $\bar{X}_F$ sont les moyennes des dards des hommes et des femmes, respectivement, qui ne sont pas tombés.

Dards non tombés en fonction du sexe



Graphique


Simulation de différents estimateurs de la moyenne
Variable 2
11.0
10.5
10.0
9.5
9.5
10.0
10.5
Variable 1
Bleu: tous, Noir: pas tombés, Rouge: moyenne pondérée H/F

Homme vs femme

Université de Montréal

- Comme c'est le fait d'être droitier vs gaucher qui a un impact sur la moyenne des observations plutôt que le fait d'être un homme vs une femme, la moyenne pondérée en fonction du sexe n'a rien changé au biais de l'estimateur pondéré par rapport au sexe. Tant chez les hommes que chez les femmes, il y a des droitiers et des gauchers, ceux-ci ont des moyennes différentes et des probabilités de réponse différentes.
- Ainsi, pour qu'on puisse enlever le biais potentiel dû à la non-réponse, il faut pouvoir bien comprendre ce qui cause la non-réponse et pouvoir bien modéliser pour obtenir un estimateur sans biais, ce qui n'est pas simple.
- Notez que dans ce cas-ci, lorsque la taille de l'échantillon augmente, le biais ne diminue pas : ainsi en prenant la moyenne des répondants, on obtient une estimation de plus en plus précise (l'écart-type diminue parce que n augmente) d'une valeur à côté de la cible !

Enquête nationale des ménages

Université de Montréal

- Tel que prévu, l'ENM **volontaire** a eu un taux de réponse beaucoup moins élevé que le formulaire long **obligatoire** qu'elle a remplacé. En 2006, 6% (des 20% de ménages qui ont reçu le questionnaire) n'ont pas répondu alors qu'il a été rapporté en 2013 que l'ENM de 2011...
- 32% des ménages (parmi les 30% qui ont reçu le questionnaire) n'ont pas répondu !
- Ce 32% est une moyenne de telle sorte que dans certaines communautés le pourcentage de non-réponse est encore plus élevé !
- L'ENM sert à donner un portrait global pour le Canada sur plusieurs questions, mais surtout un portrait le plus juste possible pour de petites communautés. Malheureusement, avec un aussi haut taux de non-réponse, le potentiel de biais importants dans les analyses locales est très grand de telle sorte que Statistique Canada n'a pas publié certains résultats pour de petites localités.

Enquête nationale des ménages et l'IRSST

Université de Montréal

- L'IRSST publie de nombreux indices statistiques comme on le verra dans la prochaine présentation.
- Le dénominateur de certains de ces indices était tiré du formulaire long et maintenant de l'ENM.

D'autres sources de biais

Université de Montréal

- Comparaison d'une nouvelle technique de déplacement de boîtes à la technique présentement recommandée.
- On demande des volontaires pour essayer la nouvelle technique ; ce sera notre groupe expérimental.
- On prend un échantillon parmi les autres qui utilisera la technique présentement recommandée et représentera le groupe contrôle.
- S'il y a une différence à la fin, pourra-t-on l'attribuer à la différence entre les deux traitements ou à la différence présente au départ entre ceux qui se sont portés volontaires pour essayer une nouvelle technique et les autres ?
- Il y a donc un biais potentiel.
- Et contrairement à la dispersion, nous ne sommes pas du tout en mesure de mesurer ce biais.

Biais : conclusion

Message : Le biais est une caractéristique externe, c'est-à-dire qu'on ne peut pas estimer le biais à partir des données que nous avons. Il faut donc planifier pour minimiser les chances qu'il y ait un biais. Par exemple, la randomisation des participants d'une étude à l'un des deux groupes d'une étude permet d'éviter un biais.

Table des matières

Exemple de recherche actuelle

Université de Montréal

Le bootstrap est une des plus importantes innovations du 20^{eme} siècle en statistique et a été inventé en 1979 par Brad Efron de l'Université Stanford.

FIGURE : Brad Efron

La méthode vise à estimer la distribution d'un estimateur ou une caractéristique de sa distribution comme l'écart-type. Elle consiste à utiliser les données pour estimer le modèle et à simuler un grand nombre de données bootstrap à partir de celui-ci. On calcule l'estimateur sur chacun d'entre eux et la distribution expérimentale de ces estimés bootstrap est l'estimation de la distribution de l'estimateur.

Données de Statistique Canada ou de l'ISQ

Université de Montréal

Considérons l'exemple simple suivant : nous tirons un échantillon aléatoire simple sans remise de taille 5 d'une population de taille 50.

Id i	Obs y_i	Poids w_i
2	9	10
17	11	10
18	18	10
37	7	10
48	14	10

L'estimateur du total est $\hat{t} = \sum_{i \in s} w_i y_i$ où s est l'échantillon.

Poids bootstrap pour estimer la dispersion

Université de Montréal

Dans le but d'estimer la dispersion d'estimateurs calculés sur des données d'échantillonnage, Rao, Wu et Yue ont introduit une méthode bootstrap basée sur les poids en 1992. Jon Rao est un professeur émérite de l'Université de Carleton toujours actif alors que Jeff Wu était un professeur à l'Université de Waterloo lorsque cette recherche a été effectuée.

(a) Jon Rao

(b) Jeff Wu

Fichier avec poids bootstrap

Université de Montréal

Pour l'exemple précédent, voici de quoi pourrait avoir l'air le fichier.

Id	Obs	Poids	Poids bootstrap 1	Poids bootstrap 2	...	Poids bootstrap 1000
i	y_i	w_i	$w^*_{1,i}$	$w^*_{2,i}$	...	$w^*_{1000,i}$
2	9	10	24,23	12,37	...	24,23
17	11	10	12,37	0,51	...	0,51
18	18	10	12,37	12,37	...	0,51
37	7	10	12,37	24,23	...	24,23
48	14	10	0,51	12,37	...	12,37

L'estimateur bootstrap de la dispersion de l'estimateur est la dispersion des 1000 estimateurs bootstrap $t^*_i = \sum_{j \in s} w^*_{i,j} y_j$.

Non-réponse

Mais dans la vraie vie, il y a de la non-réponse.

Id i	Obs y_i	Poids w_i
2	9	10
17	?	10
18	18	10
37	?	10
48	14	10

Pour faciliter la tâche des utilisateurs, on remplace généralement les valeurs manquantes par des valeurs imputées, par exemple par la moyenne des répondants (ici, $13{,}67$).

Fichier imputé

Université de Montréal

Le fichier mis à la disposition des utilisateurs devient le suivant :

Id	Obs	Poids	Poids bootstrap 1	Poids bootstrap 2	...	Poids bootstrap 1000
i	y_i	w_i	$w^*_{1,i}$	$w^*_{2,i}$	...	$w^*_{1000,i}$
2	9	10	24,23	12,37	...	24,23
17	13,67	10	12,37	0,51	...	0,51
18	18	10	12,37	12,37	...	0,51
37	13,67	10	12,37	24,23	...	24,23
48	14	10	0,51	12,37	...	12,37

Les valeurs imputées n'apportent pas vraiment de nouvelle information de telle sorte que bien que le fichier contienne 5 données, la variabilité de l'estimateur du total basé sur les données imputées est comme celle d'un échantillon de taille 3. Ainsi l'écart-type de l'estimateur basé sur les données imputées est plus grand que si tout le monde avait répondu. *Mais remarquez que les poids bootstrap sont les mêmes !*

Bootstrap pour la non-réponse

Université de Montréal

Avec Zeinab Mashreghi, étudiante au doctorat, et David Haziza, collègue à l'Université de Montréal, nous avons démontré que l'utilisation de ces poids bootstrap peut grandement sous-estimer la variabilité de l'estimateur.

Nous avons introduit de nouvelles méthodes qui tiennent compte de la non-réponse. Elles dépendent de la méthode d'imputation.

Impact de la non-réponse : conclusion

Université de Montréal

Message : Imputer des données a un impact sur la dispersion des estimateurs. Les poids bootstrap qu'on retrouve dans les fichiers des agences statistiques ne reflètent que la variabilité inhérente à la sélection de l'échantillon, pas celle due à l'imputation des données manquantes (dont souvent on ne connait même pas la présence). De nouvelles recherches apportent des solutions à ce problème.

Conclusion

Université de Montréal

- La statistique est omniprésente.
- Les concepts qu'elle manipule semblent parfois simples, mais sont souvent très subtils.
- Les statisticiens travaillent régulièrement de concert avec d'autres chercheurs.
- Un groupe de chercheurs de la taille de celui de l'IRSST devrait définitivement compter sur au moins un statisticien !