0. Introduction

MTH2302D

S. de Montigny en collaboration avec S. Le Digabel
École Polytechnique de Montréal

A2017

(v1)

Analyse de données

- Expressions à la mode :
  - Forage de données (*Data mining*).
  - Analyse d'affaires (*Business analytics*).
  - Mégadonnées (*Big data*).
- Importance de la visualisation et de l'analyse statistique d'ensembles de données :
  - Recherche scientifique.
  - Contrôle de la qualité en ingénierie.
  - Prise de décision en entreprise.
  - Élaboration de politiques publiques.
  - Sondages.

Échantillon aléatoire

- Les méthodes statistiques ne sont justifiées que si l'on dispose d'un **échantillon aléatoire**.
- Pour savoir si un échantillon de données est aléatoire ou non, il faut se renseigner sur la méthodologie qui a été utilisée pour la collecte de ces données :
- Exemple (www.cyberpresse.ca) : "Le sondage a été réalisé du 22 au 25 août 2012 par l'entremise d'entrevues téléphoniques. L'échantillon a été tiré aléatoirement dans la région de Québec, constituée de 12 circonscriptions provinciales définies selon la nouvelle carte électorale de 2011. Les répondants ont été sélectionnés de façon aléatoire parmi les citoyens de 18 ans et plus dans les ménages contactés. Au total, 1007 entrevues ont été réalisées. Les données d'ensemble ont été pondérées sur la base du recensement de 2011 en fonction du sexe, de l'âge et de la répartition démographique (circonscription) de la population, de façon à rendre les résultats conformes à la situation générale de la population de la région de Québec. Les résultats d'ensemble comportent une marge d'erreur échantillonnale de $\pm$ 3,1%, selon un intervalle de confiance de 95%. "

Échantillon aléatoire : Exemple

- Une firme de sondages contacte des gens choisis au hasard dans des listes téléphoniques. Les répondants forment un échantillon aléatoire.
- Les visiteurs d'un site web qui choisissent de répondre à un sondage proposé sur le site ne forment pas un échantillon aléatoire.
- La population d'un pays qui répond à un recensement commandé par son gouvernement ne forme pas un échantillon aléatoire.

Inférence statistique : Résumé

- On veut confirmer ou infirmer une hypothèse au sujet d'une population.
- On prélève un échantillon aléatoire dans la population.
- On utilise des règles issues de la théorie des probabilités pour évaluer l'hypothèse en se basant sur l'échantillon.
- Les règles fournissent le niveau de confiance de la conclusion obtenue.

Inférence statistique : Résumé

- Dans le cadre du projet de session, vous devrez trouver des données qui vous intéressent et les étudier à l'aide des méthodes statistiques vues en classe.
- Où trouver des données ?
  - Base de données de Statistique Canada : http://www5.statcan.gc.ca/cansim.
  - Bases de données de l'ONU : http://unstats.un.org/unsd/databases.htm.
  - etc.
- Quels types de données ?
  Univariées (histogrammes), bivariées (nuages de points), séries temporelles, etc.

Probabilités

La première partie du cours porte sur la théorie de la probabilité. Nous n'en discutons pas tout de suite, mais donnons simplement un exemple célèbre et contre-intuitif :

Paradoxe des anniversaires : Dans un groupe de n personnes, le probabilité d'avoir au moins deux personnes nées le même jour est de

$$1 - \frac{365!}{(365-n)!365^n}$$

(en ignorant les gens nés le 29 février). Cette probabilité dépasse 50% pour $n = 23$ et vaut environ 97% pour $n = 50$.

Problème de Monty-Hall : Wikipedia, Jeu interactif.