Introduction et motivation du cours

Statistiques appliquées à la gestion

Ana Karina Fermin

Université Paris-Ouest-Nanterre-La Défense

http://fermin.perso.math.cnrs.fr/

La statistique dans les formations !

Introduction à l'exploration statistique des données marketing et ventes

Alors que le volume de données provenant du marketing et des ventes, disponibles pour l'analyse croit de plus en plus rapidement, nombreux sont les gestionnaires qui se plaignent de la difficulté de tirer profit de cette abondance et de réaliser des prévisions. L'augmentation de l'incertitude (économique, réglementaire ou fiscale) les incite parfois à penser que l'on ne peut plus faire de prévisions. Fort heureusement, les techniques statistiques permettent d'extraire de la connaissance des données disponibles et d'encadrer l'incertitude pour mieux décider.

Cette formation propose aux professionnels de découvrir (ou de redécouvrir) les concepts-clés de l'analyse statistique dans le domaine marketing et ventes, et de les mettre en pratique avec Excel, R et Tableau.

La statistique dans la presse !

Chef économiste de Google : Hal Varian

LE FIGARO.fr

LE FLASH ECO

ECONOMIE > CONJONCTURE (ARCHIVES)

Le métier de statisticien sera le plus «sexy» de la décennie

Par Jean-Pierre Robin | Mis à jour le 18/10/2010 à 11:28 / Publié le 17/10/2010 à 11:27

La Toile et les moteurs de recherche sont d'extraordinaires pourvoyeurs d'informations et de statistiques. Le tout est de savoir s'en servir.

La statistique dans la presse !

Le taux de chômage peut-il s'expliquer par la qualité de l'éducation ?

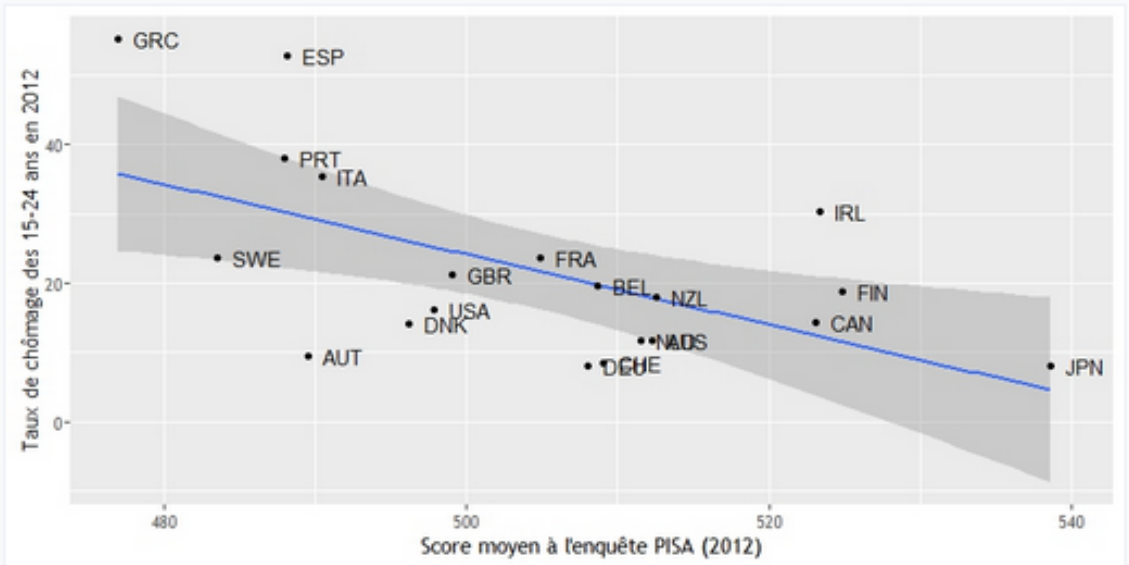
LE MONDE | 29.08.2016 à 14h07 • Mis à jour le 29.08.2016 à 14h20 |

Par Romain Damian

Plus d'écoles, moins de chômage ? C'est ce que semble affirmer une étude publiée récemment par la banque Natixis. A l'approche de la rentrée, l'économiste Patrick Artus, dans son « Flash économie » du 3 août 2016 (un papier non académique destiné aux clients de la banque), affirme que « *la qualité du système éducatif et du système de formation professionnelle joue […] un rôle majeur pour expliquer la performance économique et sociale* » d'un pays.

La statistique dans la presse !

Corrélation ne veut pas dire causalité



On peut aller plus loin que la corrélation simple grâce à une modélisation. Cette dernière, plus complexe, exige de formuler des hypothèses plus simples, mais permettrait d'estimer au mieux l'impact réel de la qualité d'éducation sur la vitalité économique d'un pays en écartant au maximum les effets dus à d'autres facteurs (les politiques monétaires ou d'austérité, par exemple).

Corrélation et causalité? Synonymes? NON!

Exemple : La consommation moyenne de chocolat par habitant est corrélée au nombre de lauréats du prix Nobel, d'après une étude de l'Américain Franz Messerli publiée en 2012.

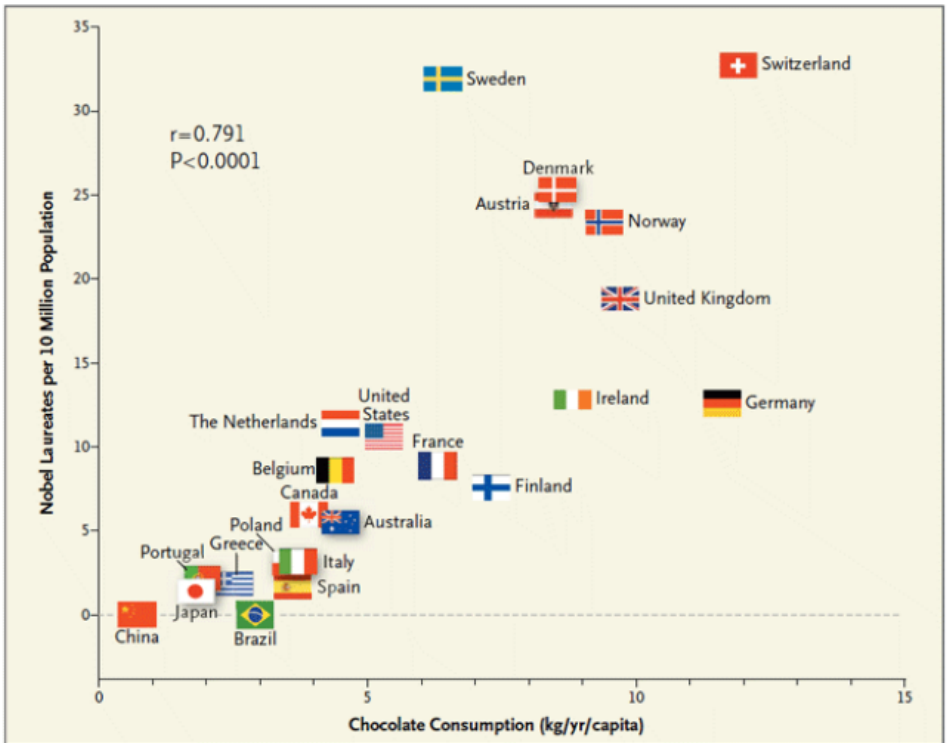
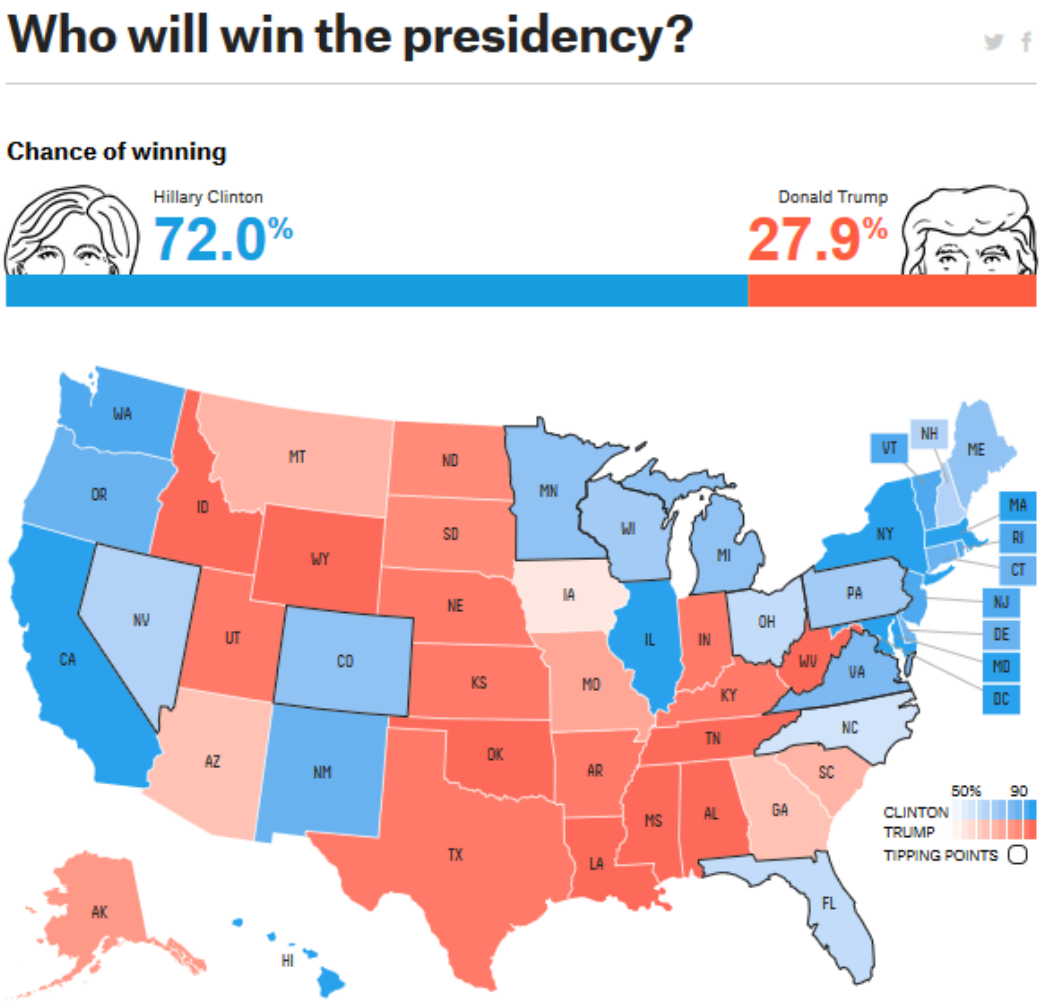


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

Nombre de prix Nobel par dix millions d'habitants en fonction de la consommation nationale de chocolat en kilogrammes par personne et par an.

Image : Franz H. Messerli, *The New England Journal of Medicine* **367**(16) (2012), p. 1562-1564


Who will win the presidency?
Chance of winning
Hillary Clinton
72.0%
Donald Trump
27.9%
WA
OR
CA
ID
NV
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
OH
KY
TN
MS
AL
WV
GA
FL
SC
NC
VA
PA
NY
VT
NH
ME
MA
RI
CT
NJ
DE
MD
DC
AK
HI
50%
90
CLINTON
TRUMP
TIPPING POINTS

Démarche à suivre

En statistique, comme dans la vraie vie, on se pose des questions, et on essaie d'y répondre. Le statisticien cherche à modéliser...

1. Visualiser les données et comprendre le problème métier.
2. Traduire le problème métier en un problème statistique.
   - Proposer une modélisation mathématique de l'expérience générant ses données.
   - Utiliser une méthode statistique pour proposer une réponse (régression, anova...).
   - Utiliser des outils statistiques pour donner des garanties sur les résultats (intervalles de confiance, tests...).
3. Utiliser les résultats pour répondre au problème métier en prenant en compte l'incertitude.

Thèmes abordés dans ce cours

- Rappels (menu du jour)
- Test de comparaison de deux distributions (chapitre 1)
  - Test de comparaison de moyennes
  - Test de proportions
- Liaison entre deux variables (chapitre 2)
- Régression (chapitre 3)
- ANOVA (chapitre 4)

Données

Les données proviennent d'une ou plusieurs variables qui sont mesurés simultanément sur un individu. Cet individu appartient à une population de taille généralement inconnue.
On dispose d'un ensemble d'observations de taille n

Exemple

- Population : Étudiants de L3 Eco Gestion
- Variable, notée X : Note de Statistiques en L2
- Données : $D_n = \{x_1, x_2, \ldots, x_n\}$

Exemple

- Population : Étudiants de L3 Eco Gestion
- Variables : Série du baccalauréat (X_1), Age (X_2), Sexe (X_3), Type de licence (X_4), Mention en L2 licence (X_5), Durée du trajet domicile-université (X_6).
- Données :

$$D_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$$

avec $\mathbf{x}_i = (x_{i1}, \ldots, x_{i6})$ le i-ème individu $(i = 1, \ldots, n)$.

Les données

Individu	X_1	...	X_j	...	X_p
1	x_{11}	...	x_{1j}	...	x_{1p}
2	x_{21}	...	x_{2j}	...	x_{2p}
...	...	...	...	...	...
i	x_{i1}	...	x_{ij}	...	x_{ip}
...	...	...	...	...	...
n	x_{n1}	...	x_{nj}	...	x_{np}

Les données "histoire de vie"

Extrait de l'enquête **histoire de vie** réalisée par l'INSEE en 2003.
La base des donnée, disponibles dans le logiciel R, contient 2000 individus et 20 variables. Les 20 variables observées sont :

```
 [1] "id"          "age"          "sexe"          "nivetud"
 [5] "poids"       "occup"        "qualif"        "freres.soeurs"
 [9] "clso"        "relig"        "trav.imp"      "trav.satisf"
[13] "hard.rock"   "lecture.bd"   "peche.chasse"  "cuisine"
[17] "bricol"      "cinema"       "sport"         "heures.tv"
```

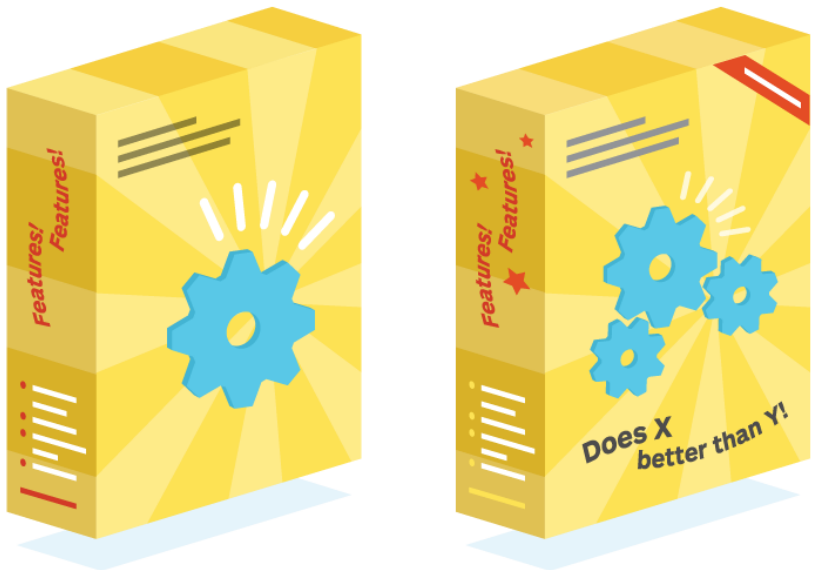
Remarque : Dans ce cours on apprendra à lire les graphiques et sorties produits à l'aide du logiciel R.

```
Variables: 20
$ id             (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14,...
$ age            (int) 28, 23, 59, 34, 71, 35, 60, 47, 20, 28, 65, 47...
$ sexe           (fctr) Femme, Femme, Homme, Homme, Femme, Femme, Fem...
$ nivetud        (fctr) Enseignement superieur y compris technique su...
$ poids          (dbl) 2634.3982, 9738.3958, 3994.1025, 5731.6615, 43...
$ occup          (fctr) Exerce une profession, Etudiant, eleve, Exerc...
$ qualif         (fctr) Employe, NA, Technicien, Technicien, Employe,...
$ freres.soeurs  (int) 8, 2, 2, 1, 0, 5, 1, 5, 4, 2, 3, 4, 1, 5, 2, 3...
$ clso           (fctr) Oui, Oui, Non, Non, Oui, Non, Oui, Non, Oui, ...
$ relig          (fctr) Ni croyance ni appartenance, Ni croyance ni a...
$ trav.imp       (fctr) Peu important, NA, Aussi important que le res...
$ trav.satisf    (fctr) Insatisfaction, NA, Equilibre, Satisfaction, ...
$ hard.rock      (fctr) Non, Non, Non, Non, Non, Non, Non, Non, Non, ...
$ lecture.bd     (fctr) Non, Non, Non, Non, Non, Non, Non, Non, Non, ...
$ peche.chasse   (fctr) Non, Non, Non, Non, Non, Non, Oui, Oui, Non, ...
$ cuisine        (fctr) Oui, Non, Non, Oui, Non, Non, Oui, Oui, Non, ...
$ bricol         (fctr) Non, Non, Non, Oui, Non, Non, Non, Oui, Non, ...
$ cinema         (fctr) Non, Oui, Non, Oui, Non, Oui, Non, Non, Oui, ...
$ sport          (fctr) Non, Oui, Oui, Oui, Non, Oui, Non, Non, Non, ...
$ heures.tv      (dbl) 0.0, 1.0, 0.0, 2.0, 3.0, 2.0, 2.9, 1.0, 2.0, 2...
```

Des données à une modélisation

Exemple : Packaging A ou packaging B

On demande à des consommateurs s'ils préfèrent, pour un produit de grande consommation qu'on veut relooker, le packaging A ou le packaging B.



A B

Des données à une modélisation

Exemple : Packaging A ou packaging B

- On demande à des consommateurs s'ils préfèrent, pour un produit de grande consommation qu'on veut relooker, le packaging A ou le packaging B.
- On interroge n personnes dans un panel de consommateurs et on inscrit les résultats dans un tableau.

Consommateur n°	1	2	3	4	5	6	. . .
Résultat	A	A	B	A	B	B	. . .

- Problème métier : choisir entre deux packaging. Choisir le packaging qui se vend le mieux !
- Idée : se baser sur des données pour prendre la décision

Quel est le travail du statisticien ?

- Donner un modèle probabiliste
- Proposer une méthodologie
- Répondre au problème métier à partir des résultats obtenus en prenant en compte l'incertitude (on ne pas sure à 100 %, il faut prendre des risques et accepter de ne pas avoir toujours raison)

Codage : on code la préférence pour A par 0 et celle pour B par 1.

Consommateur numéro	1	2	3	4	5	6	. . .
Résultat	0	0	1	0	1	1	. . .

Contexte :

- Population : Consommateurs
- Variable : réponse du consommateur, codée par une variable X qui prends de valeurs 0 ou 1.
- On note x_1, x_2, x_3, ..., les résultats successifs.
  Par exemple, le troisième consommateur interrogé préfère le packaging B ($x_3 = 1$).

Le mathématicien se place aux instants avant les interrogations d'un consommateur donné et considère celles-ci comme des expériences aléatoires : il note $X_1, X_2, X_3, \ldots$ les variables aléatoires correspondantes.

Précision mathématique du terme échantillon

Dans un étude statistique,

- on choisit au hasard les individus sur lesquels on va effectuer la mesure d'une variable
- on considère que les mesures effectuées sont les réalisations d'une variable aléatoire.

Remarque :

Dans le contexte d'une étude statistique, les résultats x_i sont alors les réalisations des variables aléatoires X_i, on les appelle les valeurs observées ou les données. Notez bien la différence entre l'utilisation des majuscules pour les variables aléatoires et les minuscules pour leurs réalisations (les données).

Imaginons qu'on tire un petit nombre de consommateurs dans un grand panel.

Exemple : Packaging A ou B

Cons. n°	Résultat
1	0
2	0
3	1
⋮	⋮

La mesure n°1 (x_1) est la réalisation d'une variable aléatoire X_1, la mesure n°2 (x_2) est la réalisation d'une variable aléatoire X_2, la mesure n°3 (x_3) est la réalisation d'une variable aléatoire X_3 etc.

Pour cet exemple, quel est la loi de chaque variable ?

Échantillon d'une loi

Les mesures sur les individus choisis au hasard sont considérées comme étant les réalisations d'une suite X_1, X_2, ..., X_n de variables aléatoires.

Definition

On dit qu'une suite X_1, X_2, ..., X_n de variables aléatoires est un échantillon d'une variable aléatoire X si :

- chaque variable aléatoire X_i a la même loi que X
- Les variables X_1, ..., X_n sont mutuellement indépendantes.

Modèle statistique paramétrique

La réponse d'un consommateur choisi au hasard est vue comme la valeur d'une variable aléatoire de loi Bernoulli.

Loi Bernoulli $\mathcal{B}(p)$, avec $p \in [0,1]$

On dit que la variable X suit une loi Bernoulli si

$$\mathbb{P}(X=1) = p \quad \text{et} \quad \mathbb{P}(X=0) = 1-p,$$

où p est un nombre réel compris entre 0 et 1 appelé paramètre de la loi. On note cette loi $\mathcal{B}(p)$.

Simulation de N réalisations sous R : rbinom(N,1,p)

```
> rbinom(20,1,0.4)
 [1] 0 1 1 0 1 0 1 1 0 0 1 0 1 0 0 0 0 1 1 0
```

Loi Bernoulli $\mathcal{B}(p)$, avec $p \in [0,1]$

On dit que la variable X suit une loi Bernoulli si

$$\mathbb{P}(X=1) = p \quad \text{et} \quad \mathbb{P}(X=0) = 1-p,$$

où p est un nombre réel compris entre 0 et 1 appelé paramètre de la loi. On note cette loi $\mathcal{B}(p)$ si

- Dans notre exemple, le paramètre p correspond à la proportion de consommateurs dans le grand panel en faveur de B.
- Si le panel de consommateurs est vraiment très grand, les enquêteurs n'ont pas le temps d'interroger tout le monde et ne peuvent accéder à la vraie valeur de p.
- Il est important de remarquer que ici p est inconnu.
- On a besoin de la théorie pour malgré tout, pouvoir dire des choses avec un degré de confiance raisonnable sur p.

Loi Binomiale $\mathcal{B}(n, p)$, avec n un entier et $p \in [0, 1]$

On dit que $X \sim \mathcal{B}(n, p)$ si

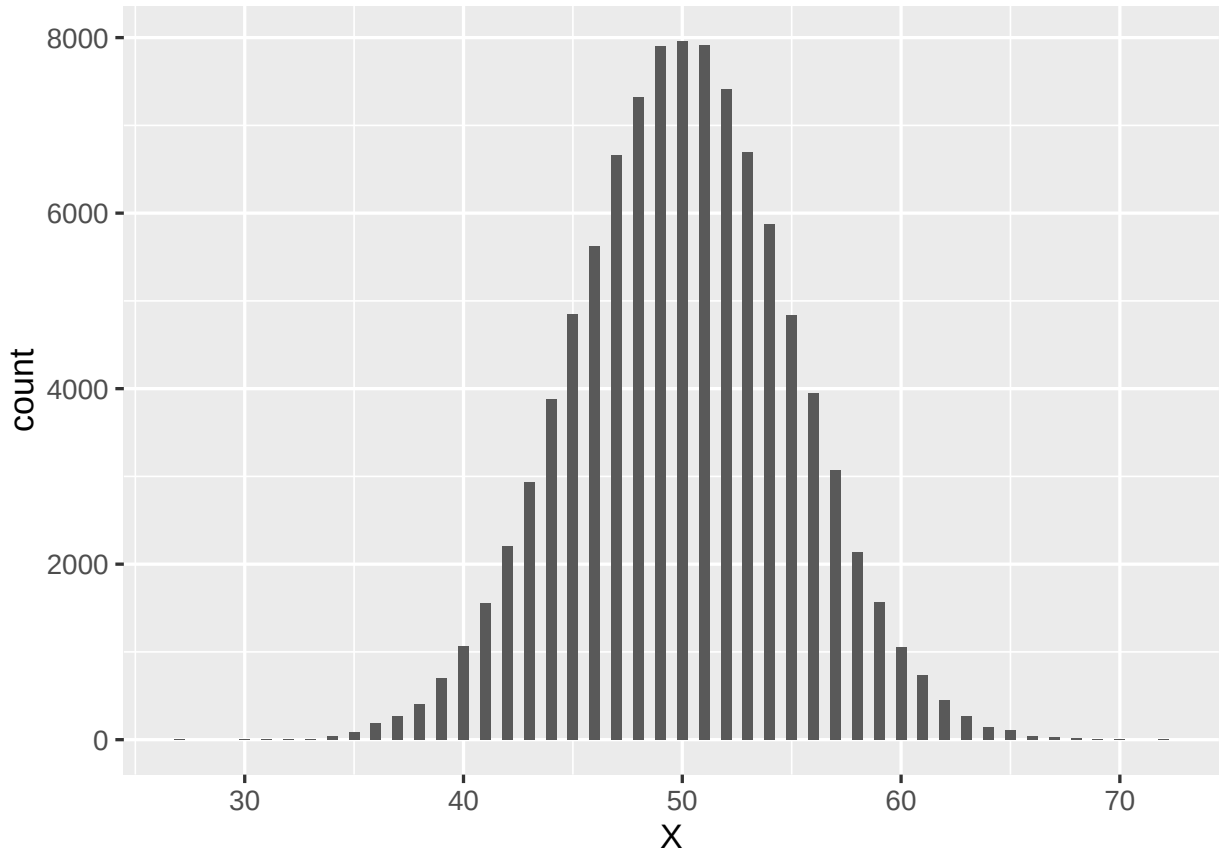
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- X modélise le nombre de fois où l'évènement B s'est produit parmi n expériences.
- X prends ses valeurs entre 0 et n

Simulation de N réalisations sous R : rbinom(N,n,p)

```
> rbinom(20,100,0.4)
 [1] 36 38 42 42 37 47 43 32 46 39 39 39 45 43 42 36 43 42 48 42
> rbinom(20,100,0.8)
 [1] 78 85 81 87 78 83 85 79 86 78 79 82 80 80 85 71 78 76 77 81
```

Simulation de 100000 réalisations d'une variable $\mathcal{B}(100, 0.5)$ sous R.



Remarque : une Binomiale $\mathcal{B}(n, p)$ s'approche à une normale (lorsque n est grand et p n'est pas trop petit).

Exercice : supposons qu'en 100 consommateurs sondés, on ait eu 58 votes en faveur de A et 42 en faveur de B. À partir de cet échantillon,

1. que peut-on déduire sur la valeur de p ?
2. peut on conclure "statistiquement" que le packaging A est préféré au packaging B ?

Estimation d'une proportion

- Estimateur ponctuelle de p : $\bar{X}_n$.
- Exemple : $n = 100$ consommateurs sondés, 58 votes en faveur de A et 42 en faveur de B.
  - A partir de notre échantillon on obtient

$$\widehat{p}_{100} = \bar{x}_{100} = 42/100 = 0.42$$

  Quelle confiance accorder à cette estimation ?
  - On fait donc appel à la théorie d'estimation d'une proportion par intervalle de confiance (on verra plus tard).

Remarque (une convention utile pour la suite) : on mettra un petit chapeau à toutes les quantités qui ne dépendent que des observations. Notez bien la différence entre $\bar{X}_{100}$ et $\bar{x}_{100}$!

Test d'hypothèses

Se poser la question si le packaging A est préférable au packaging B revient à se demander si $p = 1/2$ ou $p < 1/2$.

- Un test d'hypothèse est une démarche consistant à confronter sur un échantillon **deux** affirmations s'excluant mutuellement et portant sur la distribution de variables, appelées hypothèses statistiques.
- Les deux hypothèses statistiques confrontées n'ont pas le même statut. L'une est appelée hypothèse nulle, que l'on notera H_0, et l'autre est appelée hypothèse alternative, que l'on notera H_1.
- Choisir laquelle des deux hypothèses statistiques sera H_0 et laquelle des deux hypothèses statistiques sera H_1 n'est pas neutre. Ce choix conditionne en effet la démarche statistique.
- L'hypothèse nulle sera toujours l'hypothèse d'égalité dans ce cours.

On a besoin de la théorie de test d'hypothèses (lire les notes du cours Stat L2 de Bernard Desgraupes si vous avez besoin).

Moyenne empirique

Definition

Soit $X_1, \ldots, X_n$ un échantillon d'une loi X.

On appelle moyenne empirique la variable aléatoire $\bar{X}_n$ définie par :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

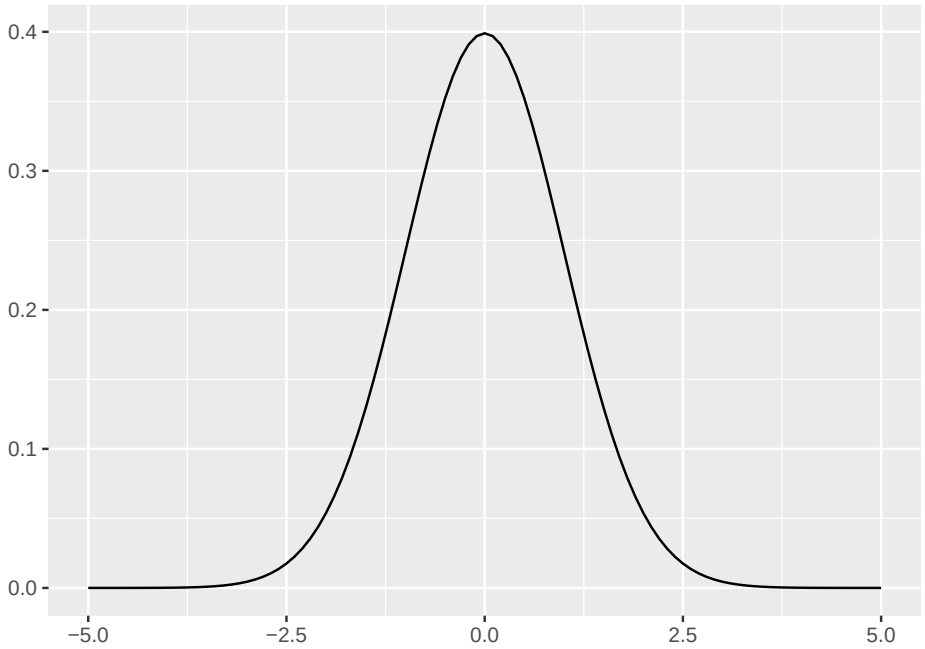
La fréquence empirique

- Dans le cas particulier d'une expérience de Bernoulli, la moyenne empirique correspond à la fréquence empirique (noté F_n plutôt que $\bar{X}_n$ dans le cours de B. Desgraupes). Dans ce cours on utilisera plutôt la notation $\bar{X}_n$.
- Puisque les valeurs de l'échantillon prennent la valeur 0 ou 1, leur somme est le nombre de fois où la valeur est 1. En divisant par n, on obtient donc la proportion des variables qui prennent la valeur 1.
- Supposons que $X \sim \mathcal{B}(p)$. On prend la fréquence empirique comme estimateur du paramètre p.

Le théorème centrale limite (TCL)

TCL

Si $X_1, \ldots, X_n$ est une suite de variables aléatoires réelles indépendantes et identiquement distribuées, alors la loi de probabilité de la quantité $\frac{(\bar{X}_n - \mathbb{E}(\bar{X}_n))}{\sqrt{Var(\bar{X}_n)}}$ se rapproche de la loi normale $\mathcal{N}(0,1)$ lorsque n es suffisamment grand.



Supposons que $X \sim B(p)$ et $X_i \sim \mathcal{B}(p)$, $i = 1, \ldots, n$

$$\mathbb{P}(X = 1) = p \quad \text{et} \quad P(X = 0) = 1 - p.$$

- Espérance de la variable X

$$\mathbb{E}(X) = 1 \times p + 0 \times (1 - p) = p$$

- Variance de la variable X

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p(1 - p)$$

- Espérance de la variable $\bar{X}_n$

$$\mathbb{E}(\bar{X}_n) = p$$

- Variance de la variable $\bar{X}_n$

$$Var(\bar{X}_n) = \frac{1}{n} p(1 - p)$$

Application du TLC :

Si l'on prend $X_i \sim \mathcal{B}(p)$, $i = 1, \ldots, n$ on retrouve qu'une Binomiale $\mathcal{B}(n, p)$ approche à une normale (lorsque n est grand).

$\frac{(\bar{X}_n - p)}{\sqrt{\frac{p(1-p)}{n}}}$ se rapproche de la loi normale $\mathcal{N}(0, 1)$

Estimation d'une proportion par intervalle de confiance

Objectif

Étant donnée une valeur α, déterminer un intervalle dépendant d'un échantillon $X_1, \ldots, X_n$ de X tel que

$$\mathbb{P}(p \in \mathrm{IC}(X_1, \ldots, X_n)) = 1 - \alpha$$

Intervalle de confiance asymptotique

Pour une valeur α fixée,

$$\mathrm{IC}(X_1, \ldots, X_n) = \left[\bar{X}_n \pm Q_Z(1 - \frac{\alpha}{2}) \sqrt{\frac{\bar{X}_n(1 - \bar{X}_n)}{n}} \right]$$

où Z est une v.a. suivant une loi normale centrée réduite.

Estimation d'une proportion par intervalle de confiance

Dans notre exemple, pour $\alpha = 0.05$, on obtient

- A l'aide d'une calculatrice et la table de la loi normale, on obtient $\sqrt{0.42 * 0.58} \approx 0.49$ et $Q_Z(0.975) = 1.96$

$$\left[0.42 - 1.96 \times \frac{0.49}{10}, 0.42 + 1.96 \times \frac{0.49}{10}\right] \approx [0.42 - 0.10, 0.42 + 0.10]$$
$$= [0.32, 0.52]$$

- A l'aide du logiciel R

```
> 0.42-qnorm(0.975)*sqrt(0.42*0.58)/sqrt(100)
 [1] 0.3232643
> 0.42 + qnorm(0.975)*sqrt(0.42*0.58)/sqrt(100)
 [1] 0.5167357
```

Réalisation de l'intervalle $\mathrm{IC}(X_1, \ldots, X_{100})$:

$$[0.3232643, 0.5167357] \approx [0.32, 0.52]$$

On peut également utiliser la fonction **binom.confit** de la libraire(binom) du logiciel R, par exemple

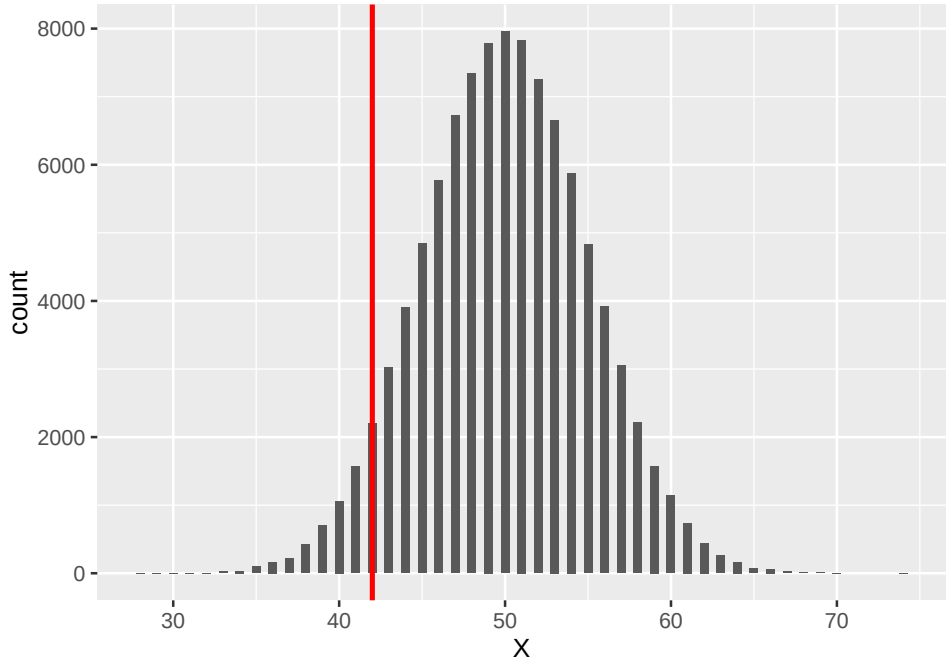
- Avec la loi exacte de $\mathcal{B}(100, 0.42)$,

```
> binom.confint(42,100,method="exact")
  method  x   n mean     lower     upper
1  exact 42 100 0.42 0.3219855 0.5228808
```

- Avec un TCL (voir le slide précédent)

```
> binom.confint(42,100,method="asymptotic")
       method  x   n mean     lower     upper
1 asymptotic 42 100 0.42 0.3232643 0.5167357
```

Test d'une proportion (motivation p-valeur)




```
> X <- rbinom(100000,100,0.5)
> mean(X <= 42)
[1] 0.06689
```

Qu'est-ce que c'est la p-value ? (motivation plus mathématique dans la prochaine séance)

Test d'une proportion (motivation p-valeur)

- Avec la loi exacte de $\mathcal{B}(100, p)$ (sous l'hypothèse que $p = 0.5$)

```
> testB=binom.test(42,100,p=0.5,alternative="less",conf.level=0.95)
> testB$estimate
probability of success
                  0.42
> testB$statistic
number of successes
                 42
> testB$p.value
[1] 0.06660531
```

- Avec l'approche à la loi normal centrée réduite (app. du TCL)

```
>testZ=z.test(42,100,p=0.5,conf.level=0.95,alternative="less")
>testZ$estimate
[1] 0.42
>testZ$statistic
[1] -1.6
>testZ$p.value
[1] 0.05479929
```

- Avec les valeurs observées $x_1, \ldots, x_{100}$ de 100 consommateurs on a construit un intervalle de confiance de p. On peut (seulement) dire avec une grande confiance que le vrai paramètre p , représentant le taux de préférence de B sur le panel, est estimé à $42\% \pm 10\%$.
- Avec les valeurs observées $x_1, \ldots, x_{100}$ de 100 consommateurs on a testé $\mathrm{H}_0 : p = 0.5$ contre $\mathrm{H}_1 : p < 0.5$ au seuil $\alpha = 5\%$. Dans cette expérience, on ne rejette pas l'hypothèse $\mathrm{H}_0 : p = 0.5$.
- Voir avec le directeur du marketing ! Tout va dépendre si le directeur de marketing est prudent et ne vaut pas se risquer au changement (à cause du coût au changement du nouveau packaging) ou si c'est le directeur de marketing qui propose ce changement en essayant de faire passer le nouveau packaging.

Bibliographie

- Notes de cours Statistique L2 Économie (Bernard Desgraupes) téléchargeables
  http://bdesgraupes.pagesperso-orange.fr/UPX/L2/index.html
- Notes de cours L3 de Gilles Stolz (Éléments de statistique pour citoyens d'aujourd'hui et managers de demain) téléchargeables
  https://studies2.hec.fr/jahia/Jahia/stoltz/pid/2949