Chapitre 2

Statistiques inférentielles

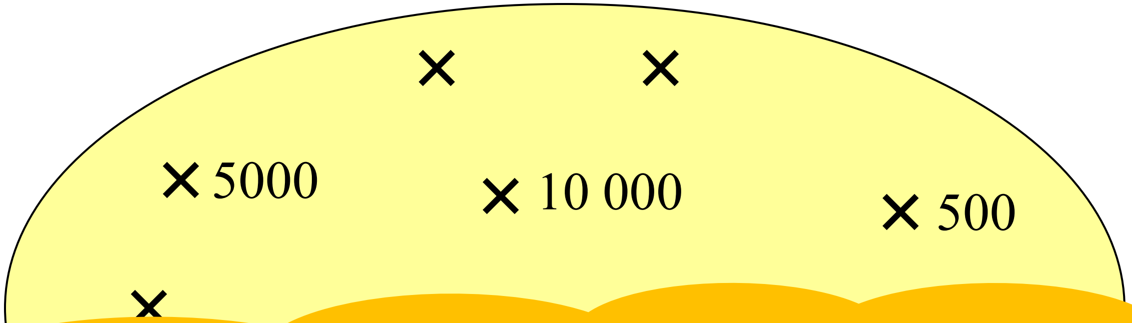
Ali JAGHDAM

ESILV - 2024

Partie 1: Estimation

UN EXEMPLE

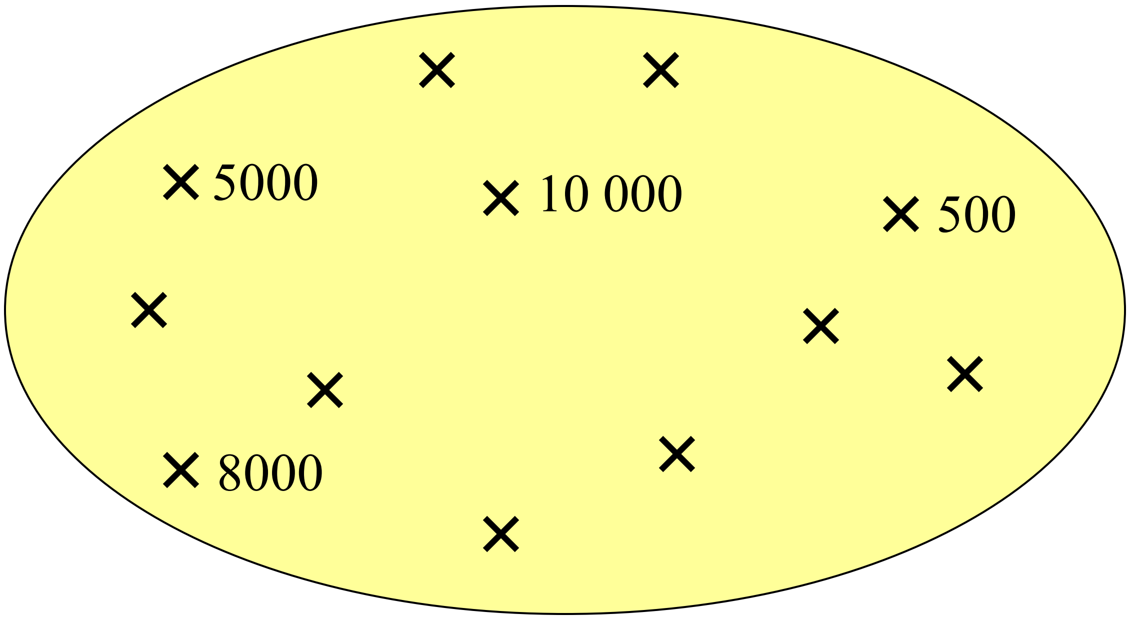
Montant quotidien des dépôts en liquide dans la banque SOCIETE GENERALE.



Comment obtenir une information sur la distribution des dépôts, sur le montant du dépôt moyen, etc....?

UNE SOLUTION SIMPLE

Observer tous les dépôts

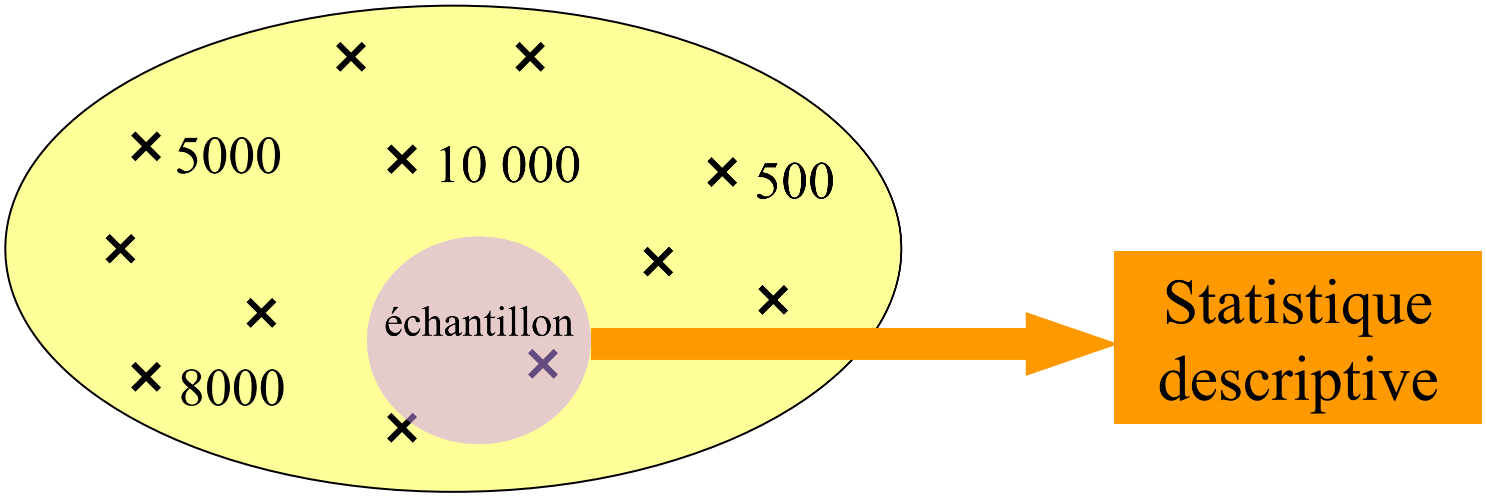


MAIS IMPOSSIBLE A METTRE EN ŒUVRE ……

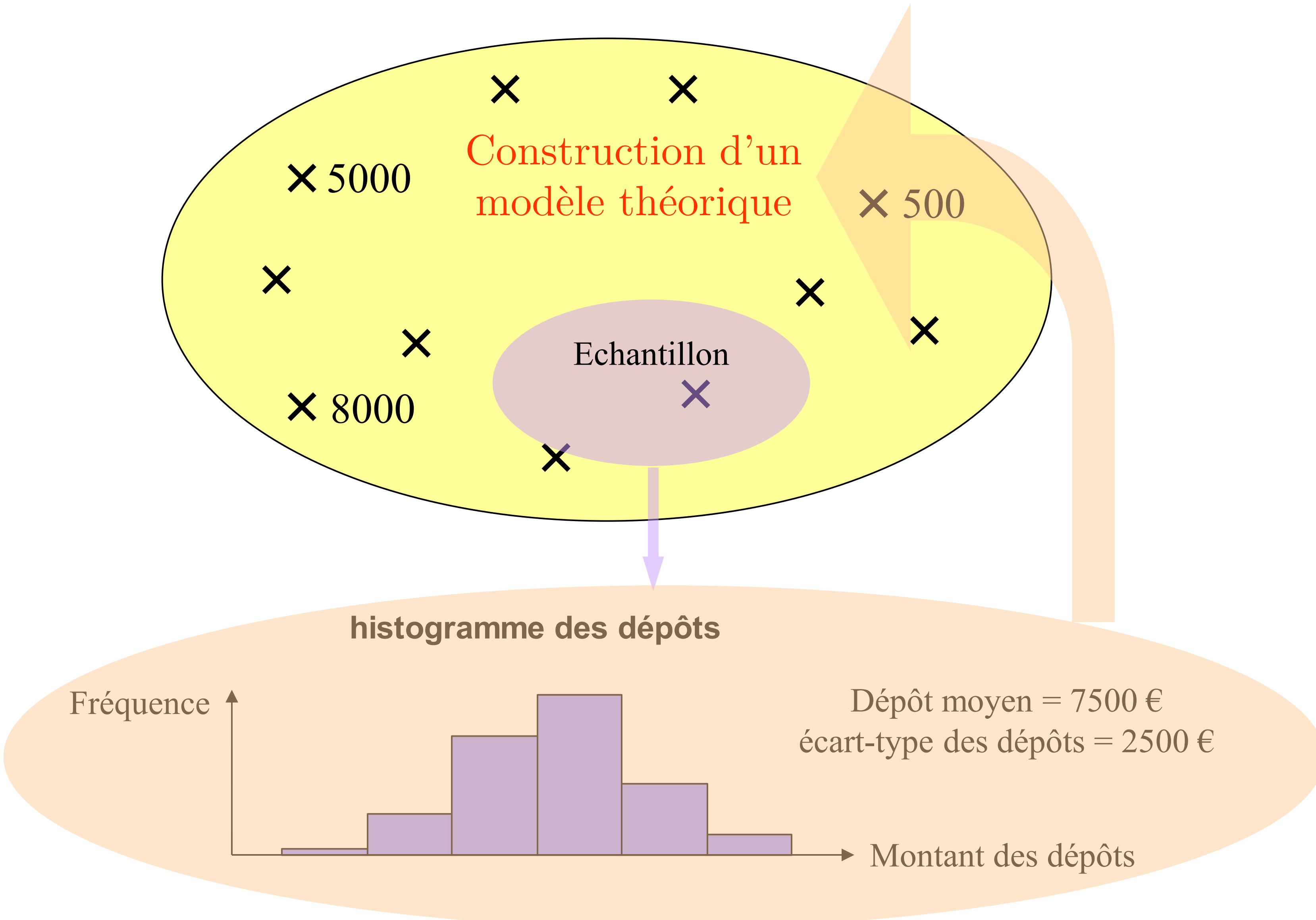
car le nombre N d’ observations est très grand, voire infini !

UNE AUTRE SOLUTION

On observe un échantillon, c.à.d. une partie de la population



INTERET DE LA STATISTIQUE DESCRIPTIVE



INTERET DE LA STATISTIQUE DESCRIPTIVE

Servir de base à la construction d'un **modèle théorique**

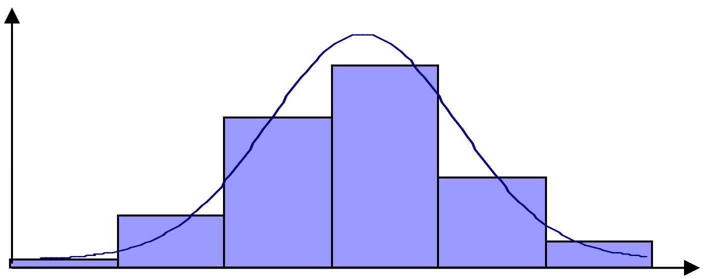
Pourquoi ?

Pour faire de la prévision:

Quelle quantité de monnaie acheter à la Banque de France en début de semaine?

Quelle quantité de liquide va-ton pouvoir faire transiter par la France ?…..

UN MODELE MATHEMATIQUE



Test d'ajustement :
On verra plus tard...

X = Montant quotidien des dépôts est une variable aléatoire

de loi Normale

de moyenne $E(X) = \mu$
de variance $V(X) = \sigma^2$

3 affirmations
à vérifier

Estimation

ESTIMATION DE LA MOYENNE μ

Comment avoir une idée sur la valeur de la moyenne μ ?

1) Prendre rendez-vous avec Irma la voyante

Problèmes: ça va me coûter cher. Puis-je lui faire confiance ?

2) Utiliser l'intuition et quelques notions de probabilités

Problème: je n'ai rien compris aux probas

C'est pas grave car on ne va plus s'amuser à jeter des dés ou tirer des cartes...

Avantage: je pourrai préciser la confiance à apporter à mon résultat

UNE METHODE INTUITIVE D'ESTIMATION DE LA MOYENNE

Pour estimer la moyenne μ inconnue de la population on utilise la moyenne $\overline{x}$ de l'échantillon.

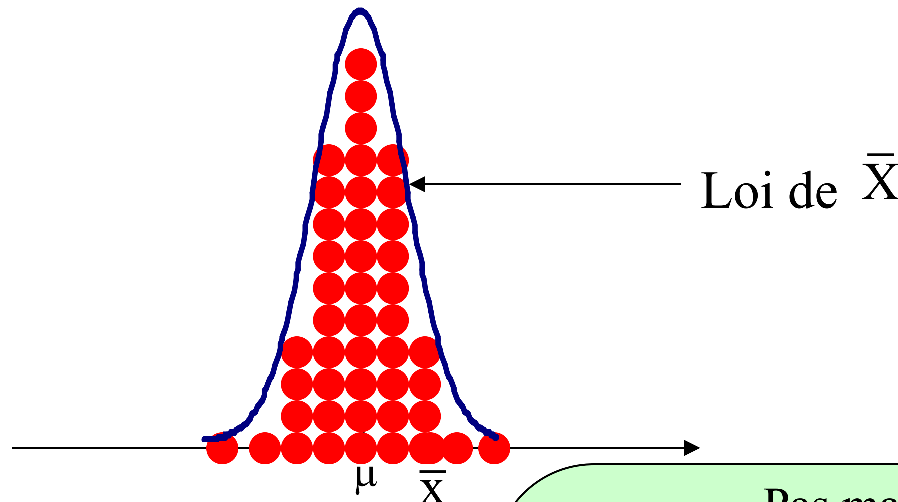
Est-on sûr de faire mieux qu'Irma ?

UNE METHODE INTUITIVE D'ESTIMATION DE LA MOYENNE

On observe n dépôts $x_1, \ldots, x_n$ sur un échantillon et on en fait la moyenne $\bar{x}$

$\bar{x}$ va-t-elle être proche de μ inconnue?

Et si j'insiste lourdement ?



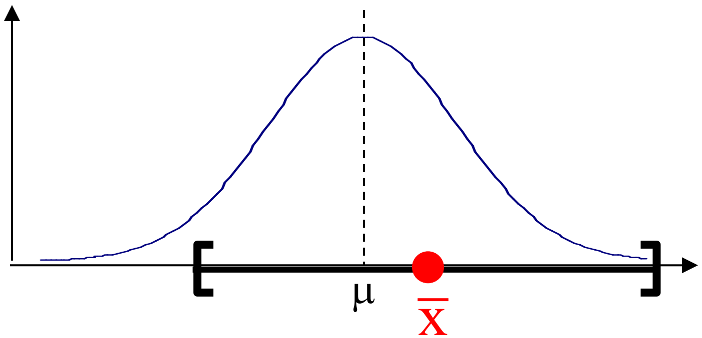
Moins bien….
Et un autre ?

Pas mal…
mais que se serait-il passé si j'avais pris un autre échantillon ?

Théorème fondamental:

Si X est une v.a. de moyenne μ et d'écart-type σ, alors la v.a. moyenne, notée $\bar{X}$, obtenue sur un échantillon de taille n tend vers une $N\left(\mu, \sigma/\sqrt{n}\right)$

Prendre une décision à partir d’un échantillon, est-ce vraiment fiable ?



Précision

Fiabilité ⟷ Probabilité

Quelle est la probabilité que la moyenne inconnue μ se trouve pas trop loin de $\overline{x}$ observée ?

Il y a $1-\alpha = 95$ chances sur 100 que l’intervalle $[\overline{x}-a\,;\,\overline{x}+a]$ contienne μ

UN PEU…….. DE PROBAS….

$\bar{X}$ suit une loi $N\left(\mu, \sigma/\sqrt{n}\right)$, donc $\sqrt{n}\left(\bar{X}-\mu\right)/\sigma$ suit une N(0,1)

$$P[-u < \sqrt{n}\left(\bar{X}-\mu\right)/\sigma < u] = 1-\alpha = 0{,}95 \xrightarrow{\text{Table}} u = \text{environ } 2 \ \ (1{,}96)$$

$$\Leftrightarrow \ P[-u\sigma/\sqrt{n} < \bar{X}-\mu < u\sigma/\sqrt{n}\ ] = 1-\alpha$$

$$\Leftrightarrow \ P[-\bar{X}-u\sigma/\sqrt{n} < -\mu < -\bar{\mathbf{X}}+u\sigma/\sqrt{n}\ ] = 1-\alpha$$

$$\Leftrightarrow \ P[\bar{X}+u\sigma/\sqrt{n} > \mu > \bar{X} - u\sigma/\sqrt{n}\ ] = 1-\alpha$$

$$\Leftrightarrow \ P[\bar{X} - u\sigma/\sqrt{n} < \mu < \bar{X}+u\sigma/\sqrt{n}\ ] = 1-\alpha$$

Il y a $1-\alpha$ chances que la moyenne inconnue μ appartienne à l'intervalle aléatoire $\left[\bar{X}-u\sigma/\sqrt{n}\,;\,\bar{X}+u\sigma/\sqrt{n}\right]$

UN PEU…….. DE PROBAS…. FIN

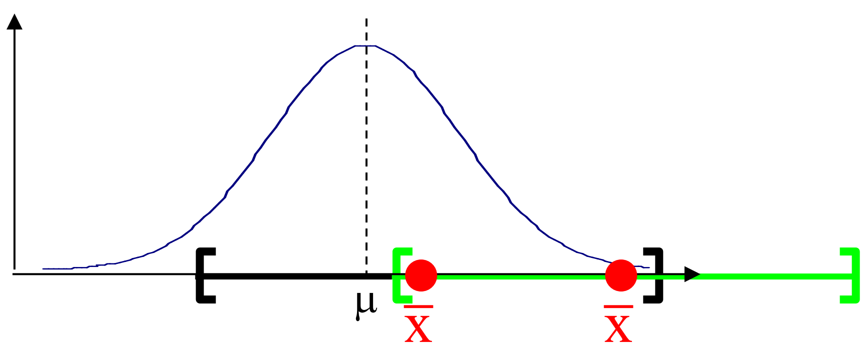
Il y a 95 chances sur 100 que la moyenne inconnue μ appartienne à l'intervalle aléatoire $\left[\bar{X} - u\sigma/\sqrt{n}\,;\,\bar{X} + u\sigma/\sqrt{n}\right]$

Une observation aléatoire nous donne la valeur $\bar{x}$ de $\bar{X}$.

$\longrightarrow \left[\bar{x} - u\sigma/\sqrt{n}\,;\,\bar{x} + u\sigma/\sqrt{n}\right]$

Une autre observation nous aurait donné une autre valeur $\bar{x}$,

……. et donc un autre intervalle …. qui ne contient pas forcément μ



Si on peut prendre une infinité d'échantillons, 95% des intervalles contiennent μ

On dira (pour simplifier) que la moyenne inconnue μ a 95 chances sur 100 d'appartenir à l'intervalle numérique $\left[\bar{x} - u\sigma/\sqrt{n}\,;\,\bar{x} + u\sigma/\sqrt{n}\right]$

Fiabilité et précision

l'intervalle $\underbrace{\left[\bar{x}-u\sigma/\sqrt{n}\,;\,\bar{x}+u\sigma/\sqrt{n}\right]}_{\text{précision}}$ a $\underbrace{1-\alpha}_{\text{fiabilité}}$ chances de contenir μ

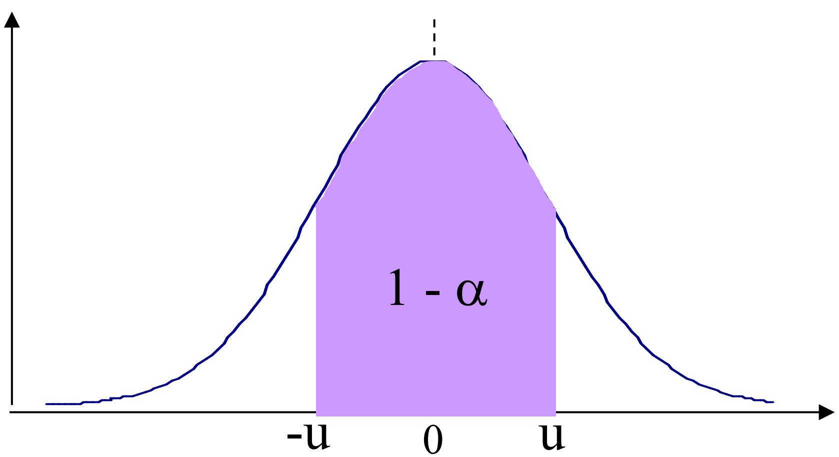
$1-\alpha$ et u sont liés par la relation $P[-u < \sqrt{n}\left(\bar{X}-\mu\right)/\sigma < u] = 1-\alpha$

$$P[-u < N(0,1) < u] = 1-\alpha$$

Fiablité augmente

u augmente

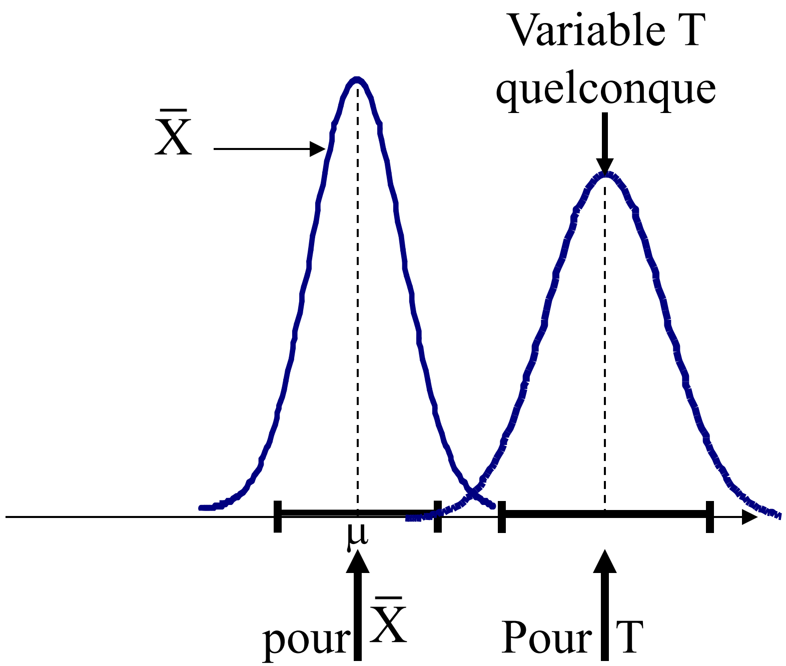
Précision diminue

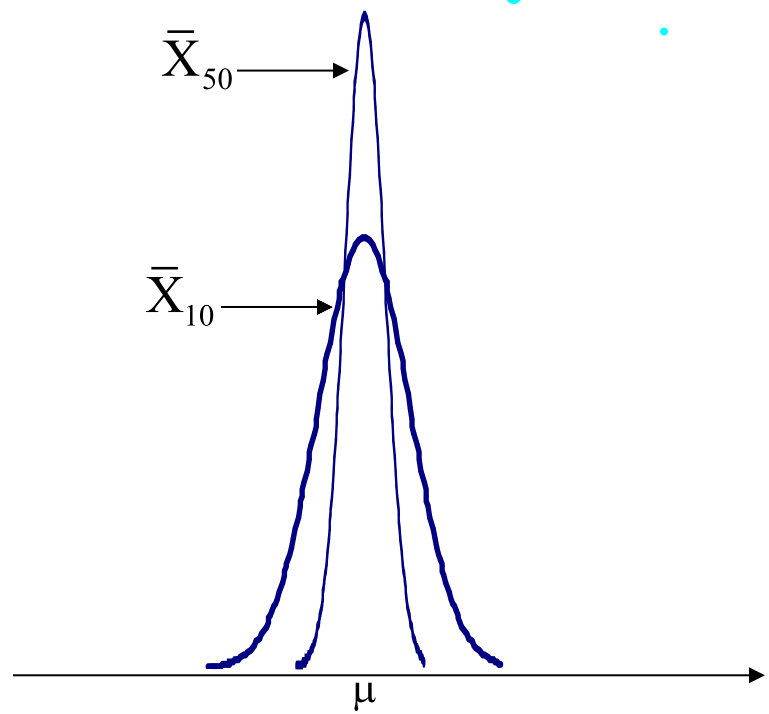


Pour quelles raisons utiliser la moyenne de l'échantillon pour estimer la moyenne de la population ?

1) Pour des raisons intuitives

2) Pour des raisons théoriques





On a une forte probabilité que l'observation soit dans cette zone

$E(\bar{X}) = \mu \qquad E(T) \neq \mu$

$\bar{X}$ est un estimateur sans biais

Plus la taille de l'échantillon grandit plus la variance diminue. Pour n infini, l'observation tombe forcément sur μ.

$\bar{X}$ est un estimateur convergent

RESUME SUR L'ESTIMATION D'UNE MOYENNE
d'une population Normale de variance connue

- Pour estimer la moyenne d'une population, on utilise la moyenne de l'échantillon $\Longrightarrow$ Estimation ponctuelle

- Pour avoir une idée de la fiabilité et de la précision du résultat on utilise un intervalle de confiance $\left[\bar{x}-u\sigma/\sqrt{n}\,;\,\bar{x}+u\sigma/\sqrt{n}\right]$ $\Longrightarrow$ Estimation par intervalle de confiance

avec u défini par $\underbrace{P[-u < N(0,1) < u]}_{\text{précision}} = \underbrace{1-\alpha}_{\text{Coefficient de confiance}}$

précision $\longleftrightarrow$ Coefficient de confiance

- Détermination de la taille d'échantillon pour une précision et un coefficient de confiance donnés

  On veut que l'intervalle soit de la forme $\bar{x} \pm \Delta$, donc $\Delta = u\,\sigma/\sqrt{n}$ et $n = (u\sigma/\Delta)^2$

ESTIMATION PONCTUELLE DE LA VARIANCE σ^2

1) Pour des raisons intuitives

Il est naturel d'estimer la variance d'une population,

par la variance $s_n^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ de l'échantillon

2) Pour des raisons théoriques

On estime la variance d'une population par la variance corrigée $s_{n-1}^2 = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$

de l'échantillon, notée aussi s^2.

Pourquoi ?

$E(S^2) = \sigma^2$ — S^2 est un estimateur sans biais de σ^2

$V(S^2) \xrightarrow[n\to\infty]{} 0$ — S^2 est un estimateur convergent de σ^2

ESTIMATION PAR INTERVALLE DE CONFIANCE DE LA VARIANCE σ² D'UNE POPULATION NORMALE

$\dfrac{(n-1)\,S^2}{\sigma^2}$ suit une loi de χ^2 à (n-1) d.d.l.

$$P\left[a < \frac{(n-1)S^2}{\sigma^2} < b\right] = 1-\alpha \quad \xrightarrow{\text{Table}} \quad \text{a et b } (> 0) \text{ pour } 1-\alpha \text{ donné}$$

$$P\left[\frac{a}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{b}{(n-1)S^2}\right] = 1-\alpha$$

$$P\left[\frac{(n-1)S^2}{a} > \sigma^2 > \frac{(n-1)S^2}{b}\right] = 1-\alpha$$

$$P\left[\frac{(n-1)S^2}{b} < \sigma^2 < \frac{(n-1)S^2}{a}\right] = 1-\alpha$$

D'où un intervalle de confiance $\left[\dfrac{(n-1)s^2}{b} ; \dfrac{(n-1)s^2}{a}\right]$

RESUME SUR L'ESTIMATION D'UNE VARIANCE

- Pour estimer la variance σ^2 d'une population, on utilise la variance corrigée s^2 de l'échantillon ⇒ Estimation ponctuelle

- Pour avoir une idée de la fiabilité et de la précision du résultat on utilise un intervalle de confiance ⇒ Estimation par intervalle de confiance

$$\left[\frac{(n-1)s^2}{b};\frac{(n-1)s^2}{a}\right]$$

avec a et b définis par

$$P\left[\chi^2_{(n-1)} < a\right] = \alpha/2$$

$$P\left[\chi^2_{(n-1)} < b\right] = 1-\alpha/2$$

ESTIMATION PONCTUELLE D'UNE PROPORTION

Dans la population il y a une proportion p d'individus possédant un certain caractère.

1) Pour des raisons intuitives

Il est naturel d'estimer la proportion p d'une population par la proportion f de l'échantillon

2) Pour des raisons théoriques

F, proportion d'échantillon, est une v.a. qui tend vers une loi $N\left(p\ ,\ \sqrt{\frac{p(1\text{-}p)}{n}}\right)$

$$E(F) = p$$

F est un estimateur sans biais de p

$$V(F) = \frac{p(1\text{-}p)}{n} \xrightarrow[n\to\infty]{} 0$$

F est un estimateur convergent de p

INTERVALLE DE CONFIANCE D'UNE PROPORTION

F tend vers une $N\left(p\ ,\ \sqrt{\dfrac{p(1\text{-}p)}{n}}\right)$, ou encore

$\sqrt{n}\,\dfrac{F-p}{\sqrt{p(1-p)}}$ est à peu près une N(0,1) dès que $n > 100$ et $0{,}1 < p < 0{,}9$

$$P\left[-u < \sqrt{n}\,\frac{F-p}{\sqrt{p(1-p)}} < u\right] = 1-\alpha \xrightarrow{\text{Table}} u \text{ pour } 1\text{-}\alpha \text{ donné}$$

$$P\left[-u\sqrt{\frac{p(1-p)}{n}} < F-p < u\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha$$

$$P\left[-F-u\sqrt{\frac{p(1-p)}{n}} < -p < -F+u\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha$$

$$P\left[F-u\sqrt{\frac{p(1-p)}{n}} < p < F+u\sqrt{\frac{p(1-p)}{n}}\right] = 1-\alpha$$

p (dans les bornes de l'intervalle aléatoire) étant inconnu, il est approché par une estimation f, d'où un intervalle de confiance

$$\left[f-u\sqrt{\frac{f(1-f)}{n}}\ ;\ f+u\sqrt{\frac{f(1-f)}{n}}\right]$$

RESUME SUR L'ESTIMATION D'UNE PROPORTION

- Pour estimer la proportion p d'une population, on utilise la proportion f de l'échantillon $\Rightarrow$ Estimation ponctuelle

- Pour avoir une idée de la fiabilité et de la précision du résultat on utilise un intervalle de confiance $\Rightarrow$ Estimation par intervalle de confiance

$$\left[f - u\sqrt{\frac{f(1\text{-}f)}{n}} ; f + u\sqrt{\frac{f(1\text{-}f)}{n}} \right]$$

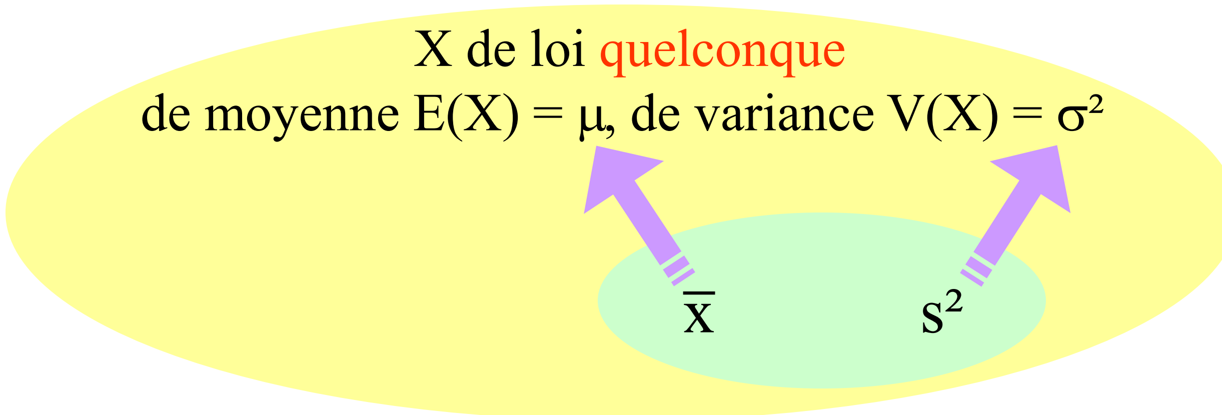
pour $n > 100$ et $0,1 < f < 0,9$
(sinon utiliser un abaque)

avec u défini par $P[-u < N(0,1) < u] = 1\text{-}\alpha$

- Détermination de la taille d'échantillon pour une précision et un coefficient de confiance donnés

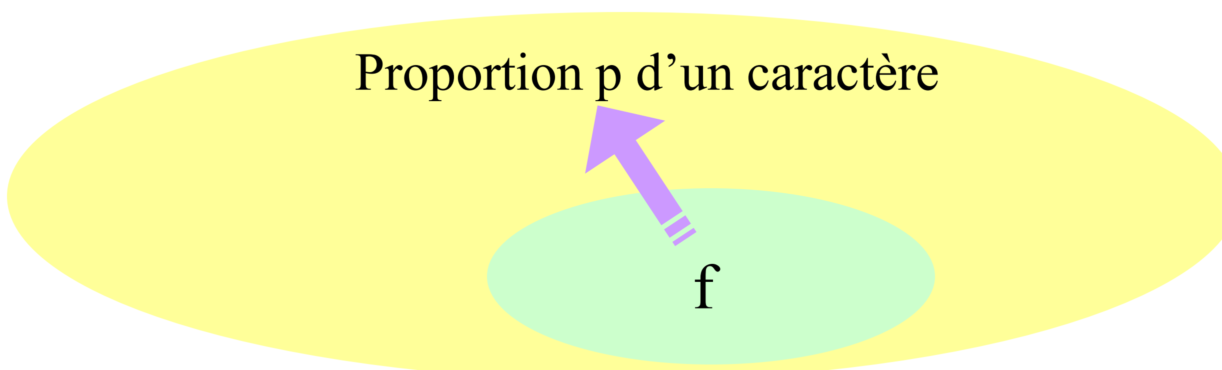
On veut que l'intervalle soit de la forme $f \pm \Delta$, donc $\Delta = u\sqrt{\frac{f(1\text{-}f)}{n}}$
et $n = \left(u/\Delta\right)^2 f(1\text{-}f)$

ESTIMATION PONCTUELLE : UNE CONCLUSION



La moyenne $\overline{x}$ de l'échantillon est une bonne estimation de la moyenne μ de la population

La variance corrigée s^2 de l'échantillon est une bonne estimation de la variance σ^2 de la population



La proportion f de l'échantillon est une bonne estimation de la proportion p de la population

INTERVALLE DE CONFIANCE D'UNE MOYENNE: EXTENSIONS

Pour obtenir un intervalle de confiance de la moyenne $\bar{x} \pm u\,\sigma/\sqrt{n}$ nous avons supposé (sans le dire) que

La taille d'échantillon est grande

La variance σ^2 est connue (ce qui en pratique est très rare)

Le taux de sondage n/N est faible (<10%)

Que peut-on faire si ces conditions ne sont pas respectées ?

INTERVALLE DE CONFIANCE D'UNE MOYENNE EXTENSIONS

	Conditions	Intervalle	u ou t défini par
σ^2 connu	Population Normale ou $n > 5$	$\bar{x} \pm u\sigma/\sqrt{n}$	$P\left[-u < N(0,1) < u\right] = 1-\alpha$
σ^2 inconnu	Population Normale ou $n > 30$	$\bar{x} \pm us/\sqrt{n}$	$P\left[-u < T_{(n-1)} < u\right] = 1-\alpha$

Loi de Student à (n-1) d.d.l.qui est approximativement N(0,1) pour n > 30

- Dans ce tableau, on suppose que l'échantillon est prélevé avec remise,
  ou
  que l'échantillon est prélevé sans remise et le taux de sondage n/N <10 %

- Dans le cas d'un échantillon prélevé sans remise, et un taux de sondage n/N >10 %, on multiplie σ ou s par le facteur d'exhaustivité $\sqrt{\dfrac{N-n}{N-1}}$

Ce correctif doit aussi être apporté pour un intervalle de confiance d'une proportion

Partie 2: Test d’hypothèses

Population

caractère observé X, de moyenne μ,
de variance σ^2

Un test consiste à

- Émettre une hypothèse, notée H_0, appelée hypothèse nulle, sur un paramètre de X, sa loi...
- Proposer une hypothèse alternative, notée H_1
- Choisir une grandeur calculée à partir de l'échantillon, appelée **statistique**
- Construire une règle de décision
- Déterminer la zone de rejet de l'hypothèse H_0 en fonction d'un risque d'erreur α que l'on veut bien accepter
- Prendre une décision

LES RISQUES D'ERREUR DANS UN TEST

La décision est

La réalité est

	Accepter H_0	Rejeter H_0
H_0 vraie	Bonne décision	Mauvaise décision: Erreur α
H_0 fausse	Mauvaise décision: Erreur β	Bonne décision

α = P(Rejeter H_0 sachant que H_0 est vraie)

β = P(Accepter H_0 sachant que H_0 est fausse)

IMPORTANCE DU CHOIX DES HYPOTHESES

H_0 est l'hypothèse à laquelle on tient le plus, la plus vraisemblable...

Il est donc plus grave de la rejeter à tort que de l'accepter à tort

Pour construire le test on se fixe α = P(Rejeter H_0 sachant que H_0 est vraie)

Souvent l'utilisateur ne calcule pas β = P(Accepter H_0 sachant que H_0 est fausse)

EXEMPLE: Les OGM sont-ils bons pour la santé?

Point de vue du consommateur

H_0 = les OGM ne sont pas bons

Point de vue de MONSANTO

H_0 = les OGM sont bons

Si β n'est pas calculé, le choix de H_0 n'est pas innocent

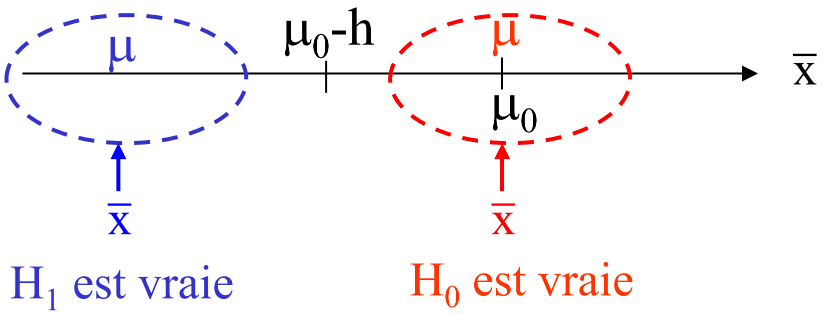
Test de comparaison d'une moyenne à une valeur donnée (variance connue)

Conditions d'application: σ connu. X suit une loi $N(\mu,\sigma)$, n quelconque
X quelconque, $n > 5$ (AFNOR)

Hypothèses: $H_0 = \{ \mu = \mu_0 \}$ contre $H_1 = \{ \mu < \mu_0 \}$

Statistique: $\bar{X}$ qui est un bon estimateur de la moyenne

Règle de décision:



Si H_0 est vraie, $\mu = \mu_0$

$\bar{x}$ est une bonne estimation de μ, donc est proche de μ

Si H_1 est vraie, $\mu < \mu_0$

$\bar{x}$ est une bonne estimation de μ, donc est proche de μ

Conclusion: Il existe μ_0-h tel que

$\bar{x} < \mu_0 - h \iff$ On rejette H_0

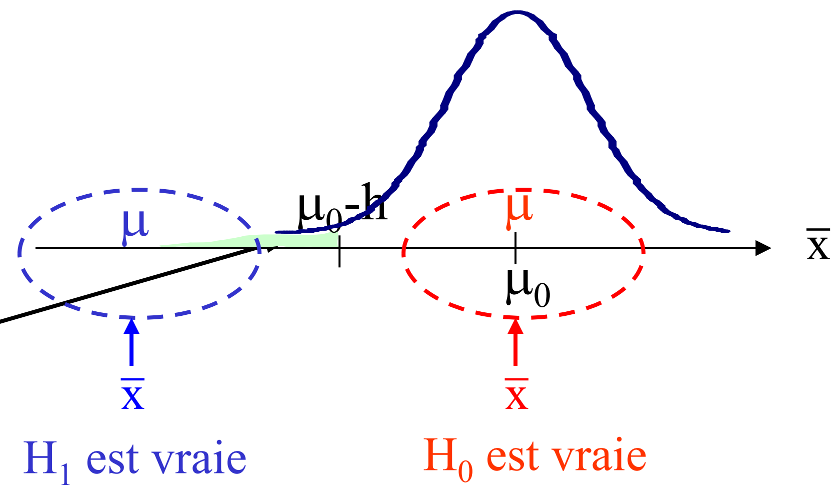
$\bar{x} > \mu_0 - h \iff$ On accepte H_0

Test de comparaison d'une moyenne à une valeur donnée (variance connue) ……. Suite 1

Règle de décision:

$\bar{x} < \mu_0\text{-}h \iff$ On rejette H_0

$\bar{x} > \mu_0\text{-}h \iff$ On accepte H_0



Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) $\Rightarrow$ $\bar{X}$ suit une $N\left(\mu_0, \sigma/\sqrt{n}\right)$

$$= P\left[\bar{X} < \mu_0 - h / \mu = \mu_0\right] = P\left[\bar{X} - \mu < \mu_0 - h - \mu / \mu = \mu_0\right]$$

$$= P\left[\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right) < \sqrt{n}\left(\frac{\mu_0 - h - \mu}{\sigma}\right) \Big/ \mu = \mu_0\right]$$

$$= P\left[N(0,1) < -\sqrt{n}\,\frac{h}{\sigma}\right]$$

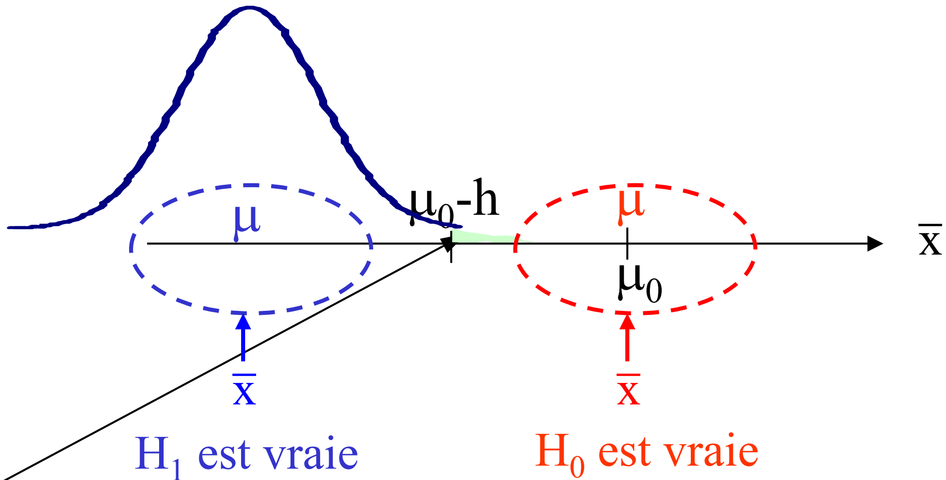
α donné $\Rightarrow$ $-\sqrt{n}\,\dfrac{h}{\sigma}$ et donc h

Test de comparaison d'une moyenne à une valeur donnée (variance connue) ……. Suite 2

Règle de décision:

$\bar{x} < \mu_0\text{-h} \iff$ On rejette H_0

$\bar{x} > \mu_0\text{-h} \iff$ On accepte H_0



Zone de rejet: permet de calculer h

Décision:

Si $\bar{x} < \mu_0\text{-h}$, on rejette H_0 avec un risque α connu de se tromper

Si $\bar{x} > \mu_0\text{-h}$, on accepte H_0 avec un risque β de se tromper

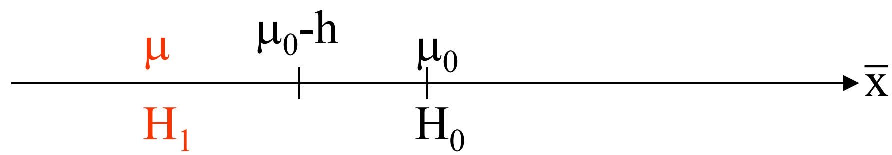
β = P(Accepter H_0 sachant que H_0 est fausse) → $\bar{X}$ suit une $N\left(\mu, \sigma/\sqrt{n}\right)$

β est fonction de μ, et n'est pas toujours calculée par l'utilisateur. Si c'est le cas, plutôt que d'accepter H_0 , il vaut mieux conclure que l'échantillon observé ne permet pas de rejeter H_0.

Test de comparaison d’une moyenne à une valeur donnée (variance connue) ……. Suite 3 ….. et fin

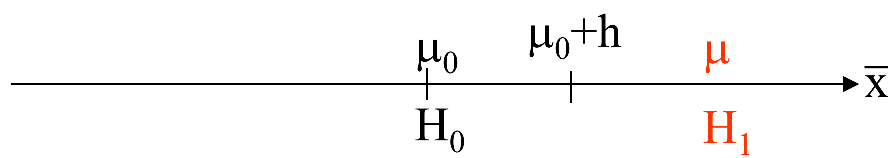
La règle de décision dépend de H_1

- $H_1 = \{\ \mu < \mu_0 \}$



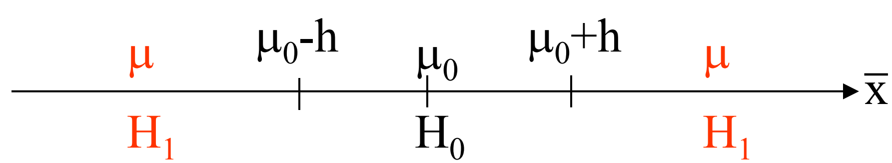
Règle: $\bar{x} < \mu_0\text{-}h \iff$ On rejette H_0

- $H_1 = \{\ \mu > \mu_0 \}$



Règle: $\bar{x} > \mu_0\text{+}h \iff$ On rejette H_0

- $H_1 = \{\ \mu \neq \mu_0 \}$



Règle: $\bar{x} < \mu_0\text{-}h$ ou $\bar{x} > \mu_0\text{+}h \iff$ On rejette H_0

Test de comparaison d'une moyenne à une valeur donnée (variance inconnue)

Conditions d'application: σ inconnu. X suit une loi $N(\mu,\sigma)$, n quelconque
X quelconque, $n \geq 30$ (AFNOR)

Hypothèses: $H_0 = \{ \mu = \mu_0 \}$ contre $H_1 = \{ \mu < \mu_0 \}$

$H_1 = \{ \mu > \mu_0 \}$

$H_1 = \{ \mu \neq \mu_0 \}$

Statistique: $\sqrt{n}\,\dfrac{\bar{X}-\mu}{S}$ suit une $T_{(n-1)}$ (approximativement $N(0,1)$ si $n \geq 30$)

Règle de décision:

$H_1 = \{ \mu < \mu_0 \}$ $\quad \bar{x} < \mu_0 - h \iff$ On rejette H_0

$H_1 = \{ \mu > \mu_0 \}$ $\quad \bar{x} > \mu_0 + h \iff$ On rejette H_0

$H_1 = \{ \mu \neq \mu_0 \}$ $\quad \bar{x} < \mu_0 - h$ ou $\bar{x} > \mu_0 + h \iff$ On rejette H_0

Zone de rejet:

$\alpha = P(\text{Rejeter } H_0 \text{ sachant que } H_0 \text{ est vraie}) = \ldots\ldots\ldots$ d'où h

Décision: en comparant $\bar{x}$ à μ_0-h ou (et) μ_0+h

Test de comparaison d'une proportion à une valeur donnée

Conditions d'application: tirage avec remise ou taux de sondage n/N < 10%
$n \geq 50$ et $np(1-p) \geq 9$ (AFNOR)

Hypothèses: $H_0 = \{ p = p_0 \}$ contre $H_1 = \{ p < p_0 \}$

$H_1 = \{ p > p_0 \}$

$H_1 = \{ p \neq p_0 \}$

Statistique: F bon estimateur de la proportion F suit une $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$

Règle de décision:

$H_1 = \{ p < p_0 \}$ $f < p_0 - h \iff$ On rejette H_0

$H_1 = \{ p > p_0 \}$ $f > p_0 + h \iff$ On rejette H_0

$H_1 = \{ p \neq p_0 \}$ $f < p_0 - h$ ou $f > p_0 + h \iff$ On rejette H_0

Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) = …….. d'où h

Décision: en comparant f à p_0-h ou (et) p_0+h

Test de comparaison de deux moyennes (variances connues)

Conditions d'application: σ_1 , σ_2 connus .

X_1 suit une $N(\mu_1,\sigma_1)$, X_2 suit une $N(\mu_2,\sigma_2)$, n_i quelconques

X_i quelconque, $n_i > 5$ (AFNOR)

Hypothèses: $H_0 = \{ \mu_1 = \mu_2 \}$ contre $H_1 = \{ \mu_1 < \mu_2 \}$

$\Updownarrow$ $\qquad H_1 = \{ \mu_1 > \mu_2 \}$

$\mu_1 - \mu_2 = 0$ $\qquad H_1 = \{ \mu_1 \neq \mu_2 \}$

Statistique: Si H_0 vraie, $\bar{X}_1 - \bar{X}_2$ suit une $N(0, \sigma_d)$ avec $\sigma_d = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

Règle de décision:

$H_1 = \{ \mu_1 < \mu_2 \}$ $\qquad \bar{x}_1 - \bar{x}_2 < -h \iff$ On rejette H_0

$H_1 = \{ \mu_1 > \mu_2 \}$ $\qquad \bar{x}_1 - \bar{x}_2 > h \iff$ On rejette H_0

$H_1 = \{ \mu_1 \neq \mu_2 \}$ $\qquad \bar{x}_1 - \bar{x}_2 < -h$ ou $\bar{x}_1 - \bar{x}_2 > h \iff$ On rejette H_0

Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) = …….. d'où h

Décision: en comparant $\bar{x}_1 - \bar{x}_2$ à h ou (et) -h

Test de comparaison de deux moyennes (variances inconnues)

Conditions d'application: n_1 et $n_2 \geq 30$ (AFNOR)

Hypothèses: $H_0 = \{ \mu_1 = \mu_2 \}$ contre $H_1 = \{ \mu_1 < \mu_2 \}$

$\Updownarrow$ $\qquad H_1 = \{ \mu_1 < \mu_2 \}$

$\mu_1 - \mu_2 = 0$ $\qquad H_1 = \{ \mu_1 \neq \mu_2 \}$

Statistique: Si H_0 vraie, $\bar{X}_1 - \bar{X}_2$ suit une $N(0, s_d)$ avec $s_d = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Règle de décision:

$H_1 = \{ \mu_1 < \mu_2 \}$ $\qquad \bar{x}_1 - \bar{x}_2 < -h \iff$ On rejette H_0

$H_1 = \{ \mu_1 > \mu_2 \}$ $\qquad \bar{x}_1 - \bar{x}_2 > h \iff$ On rejette H_0

$H_1 = \{ \mu_1 \neq \mu_2 \}$ $\qquad \bar{x}_1 - \bar{x}_2 < -h$ ou $\bar{x}_1 - \bar{x}_2 > h \iff$ On rejette H_0

Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) = …….. d'où h

Décision: en comparant $\bar{x}_1 - \bar{x}_2$ à h ou (et) -h

Test de comparaison de deux proportions

Conditions d'application: tirage avec remise ou taux de sondage n/N < 10%

$$n_1 p,\ n_1(1-p),\ n_2 p,\ n_2(1-p) \geq 5$$

Hypothèses: $H_0 = \{ p_1 = p_2 \}$ contre $H_1 = \{ p_1 < p_2 \}$

$\Updownarrow$ $\qquad\qquad\qquad H_1 = \{ p_1 > p_2 \}$

$p_1 - p_2 = 0$ $\qquad\qquad H_1 = \{ p_1 \neq p_2 \}$

Statistique: Si H_0 vraie, $F_1 - F_2$ suit une $N(0, \sigma_d)$ avec $\sigma_d = \sqrt{f_0(1-f_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$,

et $f_0 = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$ (estimation de $p = p_1 = p_2$)

Règle de décision:

$H_1 = \{ p_1 < p_2 \}$ $\qquad f_1 - f_2 < -h \iff$ On rejette H_0

$H_1 = \{ p_1 > p_2 \}$ $\qquad f_1 - f_2 > h \iff$ On rejette H_0

$H_1 = \{ p_1 \neq p_2 \}$ $\qquad f_1 - f_2 < -h$ ou $f_1 - f_2 > h \iff$ On rejette H_0

Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) = …….. d'où h

Décision: en comparant $f_1 - f_2$ à h ou (et) -h

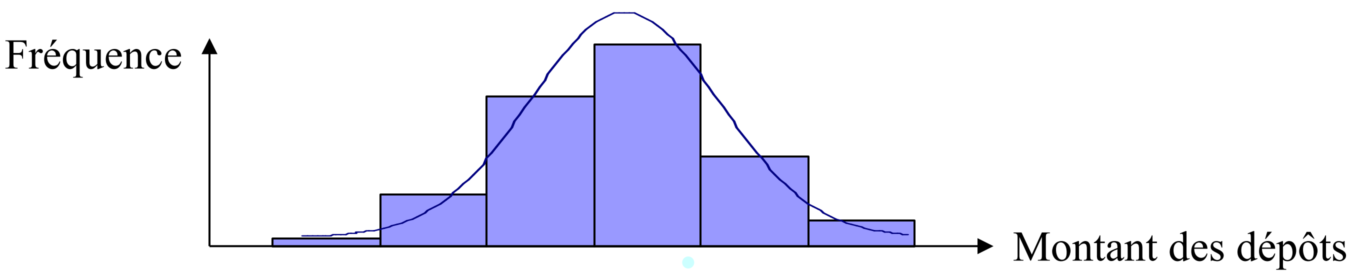
TEST D'AJUSTEMENT

UN EXEMPLE

On a observé pendant une longue période le montant hebdomadaire des dépôts en liquide dans la banque SG.

montant	[0 - 2000]	[2000 - 4000]	[4000 - 6000]	[6000 - 8000]	[8000 - 10000]	[10000 - 12000]
effectif	10	58	166	222	100	28

histogramme des dépôts



Le montant hebdomadaire des dépôts peut-il être considéré comme une loi Normale ?

Première étape: estimation des paramètres

Estimation ponctuelle: $\Rightarrow$ $\bar{x} = 6561$

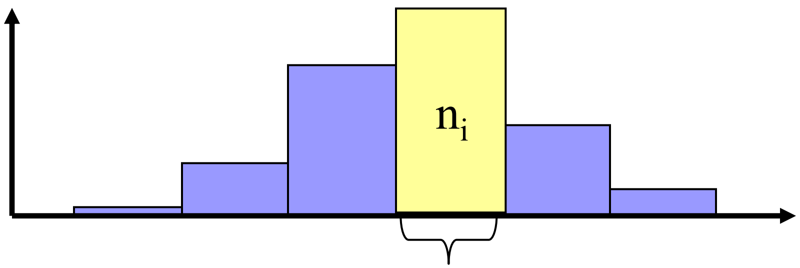
$s = 2016$

Deuxième étape: ajustement à une loi normale

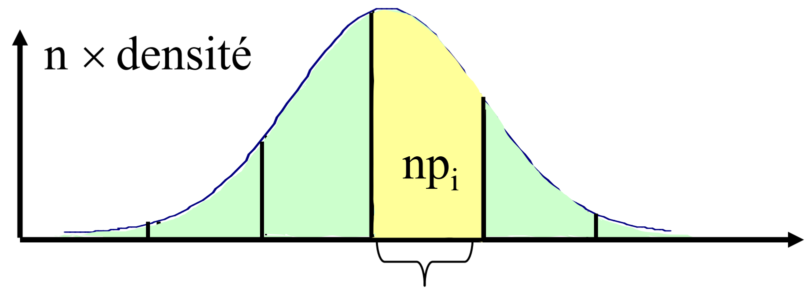
Le montant hebdomadaire des dépôts est-il issu d'une v.a. X de loi Normale ($\mu = 6561$; $\sigma = 2016$) ?

Observations

X loi N(6561 ; 2016)



Classe i



Classe i

$p_i = P(X \in \text{classe } i) = \int \phi(x)dx$

n_i = effectif observé dans la classe i
= aire de la surface de la classe i

np_i = effectif théorique dans la classe i
= aire de la surface de la classe i

Si les observations sont issues de la loi Normale (6561 ; 2016), les effectifs observés n_i dans la classe i doivent être très proches des effectifs théoriques np_i.

Comment mesurer globalement la proximité des deux graphiques ?

1) Une mesure intuitive

La proximité des 2 aires peut être mesurée par $n_i - np_i = Efo_i - Eft_i$

Effectif observé

Effectif théorique

Plus cette quantité est faible (positive ou négative) plus les aires sont proches

La proximité des 2 graphiques peut être mesurée par $\sum_i \left(Efo_i - Eft_i\right)^{2}$

Cependant, si les écarts positifs compensent les écarts négatifs, cette quantité peut être très faible avec des valeurs très différentes dans les 2 graphiques

2) Une mesure probabiliste

n_i est une observation d'une v.a.

Donc $\sum_i \left(Efo_i - Eft_i\right)^2$ est une observation d'une v.a. dont la loi n'est malheureusement pas connue. On utilise la quantité

Nombre de classes de la variable

Nombre de paramètres estimés

$$\boxed{D = \sum_i \frac{\left(Efo_i - Eft_i\right)^2}{Eft_i}}$$

qui suit une loi de χ^2 à $\nu = (k - r - 1)$ d.d.l.

Remarques importantes

Le nombre de classes et l'amplitude des classes n'a pas d'importance

L'utilisation de la loi du χ^2 n'est justifiée que si les effectifs théoriques de chacune des classes est supérieur ou égal à 5.

Si ce n'est pas le cas, il faut regrouper des classes contiguës afin d'augmenter les effectifs.

Le nombre de degrés de liberté de la loi du χ^2 dépend du nombre de classes après regroupement.

Résumé sur le test d'ajustement

Hypothèses: H_0 = { les observations sont issues d'une certaine loi}

contre H_1 = { les observations ne sont pas issues de cette loi }

Statistique: Si H_0 vraie, $D = \sum_i \frac{\left(Efo_i - Eft_i\right)^2}{Eft_i}$ est une χ^2 à $\nu = (k - r - 1)$ d.d.l.

Règle de décision:

$d > h \iff$ On rejette H_0

$d < h \iff$ l'échantillon observé ne permet pas de rejeter H_0

Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) = $P\left(\chi^2_\nu > h\right)$, d'où h

Décision: en comparant d à h

TEST D'INDEPENDANCE DE 2 VARIABLES

UN EXEMPLE

Montant des dépôts en liquide dans la banque Ibardinescroak en 2005

Catégories socio-professionelles (X \ Y)	Moins de 500 €	Entre 500 et 2000 €	Plus de 2000 €	Total
Professions libérales	20	50	180	250
Fonctionnaires	50	30	20	100
employés	230	10	10	250
Total	300	90	210	600

Y a-t-il un lien entre le montant des dépôts et la catégorie socio-professionnelle ?

	y_1	y_2	y_3	Total
x_1	20	50	180	250
x_2	50	30	20	100
x_3	230	10	10	250
Total	300	90	210	600

	…	y_j	…	Total
…	…	…	…	…
x_i	…	n_{ij}	…	$n_{i.}$
…	…	…	…	…
Total	…	$n_{.j}$	…	n

Etudions la distribution de chacune des catégories professionnelles

	y_1	y_2	y_3	Total
x_1	$\frac{20}{250}=8\%$	$\frac{50}{250}=20\%$	$\frac{180}{250}=72\%$	$\frac{250}{250}=100\%$
x_2	$\frac{50}{100}=50\%$	$\frac{30}{100}=30\%$	$\frac{20}{100}=20\%$	$\frac{100}{100}=100\%$
x_3	$\frac{230}{250}=92\%$	$\frac{10}{250}=4\%$	$\frac{10}{250}=4\%$	$\frac{250}{250}=100\%$
Total	$\frac{300}{600}=50\%$	$\frac{90}{600}=15\%$	$\frac{210}{600}=35\%$	$\frac{600}{600}=100\%$

	…	y_j	…	Total
…	…	…	…	…
x_i	…	$n_{ij}/ n_{i.}$	…	$n_{i.} / n_{i.}$
…	…	…	…	…
Total	…	$n_{.j}/ n$	…	n/ n

Si la variable X était indépendante de la variable Y, les distributions de chaque modalité de X seraient identiques, et identiques à celle du total

$n_{ij}/ n_{i.} = n_{.j}/ n$ pour tout i et j

$$\boxed{n_{ij} = \frac{n_{i.}\,n_{.j}}{n}}$$

Tableau initial

	…	y_j	…	Total
…	…	…	…	…
x_i	…	n_{ij}	…	$n_{i\cdot}$
…	…	…	…	…
Total	…	$n_{\cdot j}$	…	n

Tableau lorsque X et Y sont indépendantes

	…	y_j	…	Total
…	…	…	…	…
x_i	…	$\frac{n_{i}.n_{\cdot j}}{n}$	…	$n_{i\cdot}$
…	…	…	…	…
Total	…	$n_{\cdot j}$	…	n

Si les 2 variables X et Y sont indépendantes, les 2 tableaux doivent contenir des valeurs très proches:

n_{ij} doit être très proche de $\frac{n_{i}.n_{\cdot j}}{n}$, pour tout i et j

Comment mesurer globalement la proximité des deux tableaux ?

1) Une mesure intuitive

La proximité de 2 cellules peut être mesurée par $n_{ij} - \dfrac{n_{i}.n._{j}}{n} = Efo_{ij} - Eft_{ij}$

Effectif observé

Effectif théorique

Plus cette quantité est faible (positive ou négative) plus les cellules sont proches

La proximité des 2 tableaux peut être mesurée par $\sum_{i,j}\left(Efo_{ij} - Eft_{ij}\right)^{2}$

Cependant, si les écarts positifs compensent les écarts négatifs, cette quantité peut être très faible avec des valeurs très différentes dans les 2 tableaux

2) Une mesure probabiliste

Nombre de modalités de la variable en ligne

n_{ij} est une observation d'une v.a.

Donc $\sum_{i,j}\left(Efo_{ij} - Eft_{ij}\right)^{2}$ est une observation d'une v.a. dont la loi n'est malheureusement pas connue. On utilise la quantité

Nombre de modalités de la variable en colonne

$$\boxed{D = \sum_{i,j}\frac{\left(Efo_{ij} - Eft_{ij}\right)^{2}}{Eft_{ij}}}$$

qui suit une loi de χ^2 à $\nu = (\ell-1)(c-1)$ d.d.l.

Remarques importantes

L’utilisation de la loi du χ^2 n’est justifiée que si les effectifs théoriques de chacune des cellules est supérieur ou égal à 5.

Si ce n’est pas le cas, il faut regrouper des modalités d’une des 2 variables afin d’augmenter les effectifs.

Le nombre de degrés de liberté de la loi du χ^2 dépend du nombre de modalités des 2 variables après regroupement.

Résumé sur le test d'indépendance de deux variables

Hypothèses: H_0 = { les 2 variables X et Y sont indépendantes }

contre H_1 = { les 2 variables X et Y sont dépendantes }

Statistique: Si H_0 vraie, $D = \sum_{i,j} \frac{\left(Efo_{ij} - Eft_{ij}\right)^2}{Eft_{ij}}$ est une χ^2 à $\nu = (\ell-1)(c-1)$ d.d.l.

Règle de décision:

$d > h \iff$ On rejette H_0

$d < h \iff$ l'échantillon observé ne permet pas de rejeter H_0

Zone de rejet:

α = P(Rejeter H_0 sachant que H_0 est vraie) = $P\left(\chi^2_\nu > h\right)$, d'où h

Décision: en comparant d à h