Echantillonnage Estimation

Rappels Notions à maîtriser avant l'étude de ce chapitre

∗ **X** est une variable aléatoire mesurant un caractère λ après **N** tirages (échantillon d'effectif N).

Dans la population **m** est la moyenne de **X** et **σ** son écart type. Et donc V(X) = **σ²**.

Si N est grand

X nombre d'éléments de caractère λ X = somme de N variables de Bernouilli	$\bar{X}$ est la moyenne de X dans des échantillons de taille N	f fréquence de λ dans l'échantillon si x =$\sum X_i$ $(bernouilli)$ f = $\frac{X}{N}$
E(X)=Np σ(X) = $\sqrt{Npq}$ V(X) = Npq	E($\overline{X}$) = m (moyenne ds la population) σ($\overline{X}$) = σ($\frac{X}{N}$) = $\frac{\sigma(X)}{N} = \frac{\sigma}{\sqrt{N}}$ V($\overline{X}$)=[σ($\overline{X}$)]² = $\frac{\sigma^2}{N}$	E(f) =$\frac{E(X)}{N} = \frac{Np}{N}$ = p **σ**(f)=$\frac{\sigma(X)}{N} = \frac{\sqrt{Npq}}{N} = \sqrt{\frac{pq}{N}}$ V(f) =[σ(X)]² = $\frac{pq}{N}$

Considérations générales

Lorsqu'on veut recueillir des informations sur une population statistique, on procède par **sondage** dés lors que la méthode exhaustive ou **recensement** n'est pas appropriée (pour des raisons qui peuvent être diverses).

Pour procéder à un sondage on doit
1) prélever un échantillon aléatoire de la population c'est la phase d'**échantillonnage**.
2) déduire de cet échantillon les caractéristiques chiffrées probables de la population, c'est la phase de l'**estimation**.

Les problèmes qui se posent autour de l'échantillonnage sont **le choix de la taille** de l'échantillon et **le choix de la méthode** d'échantillonnage qui doit à la fois respecter autant que possible le hasard pur et bien sûr composer avec les contraintes contextuelles. L'échantillon est dit « **représentatif** » lorsqu'on pense que l'on peut sans risque étendre les conclusions qu'il rend à l'ensemble de la population.

L'estimation, elle, ne peut déboucher sur une certitude. Elle va formuler ses conclusions dans une phrase qui ressemble à ça : « avec un risque de **5%** on peut dire que le pourcentage de la population possédant un caractère λ se situe **entre 18% et 22%** »
Donc, une estimation produit toujours deux données chiffrées : **le risque** (**5%**), qui mesure la fiabilité du sondage et **l'intervalle de confiance** (**[18% ; 22%]**) qui traduit la précision du sondage.
Toute estimation qui ne rend pas compte du risque et des doutes qui lui sont inhérents est une mauvaise estimation.

En fait les caractéristiques de l'échantillonnage et l'utilité de l'estimation sont intimement liées.
Il est bien évident, par exemple, que l'augmentation de la taille de l'échantillon prélevé va à la fois réduire le risque et l'amplitude de l'intervalle de confiance.
Pour s'en convaincre, il suffit de remarquer que les échantillons de **N** pièces prélevés dans une population nombreuse où l'on trouve le caractère λ avec une fréquence **p** obéissent à la loi binomiale (des tirages exhaustifs dans une population nombreuse équivalent à des tirages avec remise).
Si X est le nombre d'individus possédant le caractère λ dans l'échantillon on a : **P(X=n) =** $C_N^n p^n (1-p)^{N-n}$
Par exemple si **p = 1/6 = 0,1666 = 16,66%** et qu'on prélève des échantillons de taille **10** On aura
X = 0 pour 16% des échantillons
X = 1 pour 32% des échantillons
X = 2 pour 29% des échantillons
X = 3 ou + pour 23% des échantillons
En supposant que les échantillons prélevés respectent exactement ces proportions on pourrait dire qu'il y a 61% de chances (32+29) pour que la fréquence de λ soit entre 1 sur 10 et 2 sur 10. (0,1 < P(λ) < 0.2 risque de 39%)
Si je prélevais des échantillons de taille 100, je devrais en trouver un maximum pour X = 16 et X = 17 et le produit de la précision de l'encadrement de P(λ) (0,16<P(λ) < 0,17 ici 0,01) par le risque (r%) aura diminué, ce qui signifie que l'estimation sera meilleure.

Mais en fait ce n'est pas comme cela qu'on procède : on tire un échantillon et un seul au hasard, on mesure la fréquence de λ dans cet échantillon et on en déduit que p1% ≤ P(λ) ≤ p2% avec un risque de r%.
Pour donner un exemple :

N = 12 X = 2	→ au risque de 5% on a	3% ≤ P ≤ 52%	(valeur réelle 16,6%)
N = 120 X = 20	→ au risque de 5% on a	10% ≤ P ≤ 23%	(valeur réelle 16,6%)
N = 2000 X = 333	→ au risque de 5% on a	15% ≤ P ≤ 17%	(valeur réelle 16,6%)

Méthodes d'échantillonnage

● **Tirages au hasard**
On numérote la population, on tire des nombres au hasard, on prélève l'échantillon de population correspondant aux nombres tirés.

∗ **Sondage systématique**
On détermine la taille N de l'échantillon et on prélève systématiquement 1 individu sur n jusqu'à atteindre le chiffre N .

∗ **Sondage par grappes**
On prélève des grappes d'individus localisés au même endroit ce qui diminue le coût

∗ **Sondage avec probabilités inégales**
S'il existe, au tirage, des disparités entre individus, on affecte les individus ayant la probabilité p d'être tirés d'un coefficient (un poids) 1/p

∗ **Sondage à plusieurs degrés**
Par exemple on constitue des groupes au hasard puis on tire au hasard dans chaque groupe.
Coût et précision diminués.

∗ **Méthode des quotas**
A partir d'informations antérieures, on construit un échantillon aussi représentatif que possible de la population ce qui signifie que les quotas de certains caractères jugés en corrélation avec la caractéristique étudiée sont respectés.
(Par exemple 1000 hommes 900 femmes 700 ouvriers 400 25 – 35 ans ...). C'est une méthode souple et de faible coût donc souvent utilisée mais échappant aux règles de probabilités du fait qu'elle néglige le hasard. on ne connaît ni la précision ni la fiabilité des estimations tirées de ce procédé.

∗ **Sondage stratifié**
Lorsqu'on connaît précisément certaines caractéristiques de la population on peut constituer des groupes homogènes selon ces critères (strates) et on tire indépendamment un échantillon aléatoire dans chaque strate. Gain de précision par rapport aux autres méthodes utilisant les groupements.

Distribution d'échantillonnage

Il s'agit de déterminer les caractéristiques approximatives de l'échantillon à partir d'une population connue.

Population totale d'effectif **P**
Echantillon prélevé de taille **N**.
Tirage avec remise (entre parenthèses résultats avec tirage exhaustif)

∗ Si X est le nombre d'éléments possédant un caractère λ.
Dans la population on a **E(X) = m** et **V(X) = σ^2**

On prélève un échantillon de taille N dans cette population
Moyenne de l'échantillon

$\overline{X}$ est la moyenne observée dans l'échantillon, m et σ des valeurs réelles dans la population.

∗ Soit un échantillon de taille **N** on a

$E(\overline{X}) = m$ $\quad V(\overline{X}) = \frac{\sigma^2}{N}$ $\quad (V(\overline{X}) = \frac{\sigma^2}{N}\frac{P-N}{P-1}$ si tirages exhaustifs = sans remise)

Plus la variance qui décroit avec N est faible, plus la mesure est précise.

∗ Lorsque **N** est grand (> 30)

La loi de **Y =** $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{N}}}$ est une loi **N(0,1)** $\quad$ (**Y =** $\frac{\overline{X}-m}{\sqrt{V(\overline{X})}}$ si tirages exhaustifs)

Et les relations **P(|Y| < 1,96) = 0,95** et **P(|Y| < 2,58) = 0,99** permettent de donner un encadrement de $\overline{X}$ au risque de **5%** ou de **1%.**

Par exemple de **- 1,96 <** $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{N}}}$ **< + 1,96** on tire

$$\mathbf{m - 1{,}96\,\frac{\sigma}{\sqrt{N}} < \overline{X} < m + 1{,}96\,\frac{\sigma}{\sqrt{N}}}$$

∗ Lorsque **N** est faible (< 30)
Si la loi de X est une loi normale alors la loi de $\overline{X}$ est une loi normale.
On ne peut rien dire de plus.

On connaît la moyenne et l'écart type vrais. Quel est selon la taille de l'échantillon l'intervalle de situation de la moyenne observée (f) au risque de r% ?

Dans une population la moyenne de **X** est **m =** 800 et l'écart type est **σ = 60** .
Que peut-on dire de la moyenne de **X** dans un échantillon de **100** individus (tirage exhaustif)?

$E(\overline{X})$ **= 800** $V(\overline{X}) = \frac{(60)^2}{100}$ **= 36** et **Y** $= \frac{\overline{X}-800}{\sqrt{36}}$ suit une loi **N(0,1)**

On a donc **P(|Y| < 1,96) = 0,95** (extrait des tables de la loi normale)
|Y| < 1,96 s'écrit

$$-1{,}96 < \frac{\overline{X}-800}{\sqrt{36}} < +1{,}96 \text{ ou } -1{,}96\sqrt{36} < \overline{X} - 800 < +1{,}96\sqrt{36}$$

et donc au risque de **5%** on a $800 - 1{,}96\sqrt{36} < \overline{X} < 800 + 1{,}96\sqrt{36}$
On a donc 95% de chances pour que $788{,}24 < \overline{X} < 811{,}76$

On a aussi **P(|Y| < 2,58) = 0,99** (extrait des tables de la loi normale)
et donc au risque de **1%** on a $800 - 2{,}58\sqrt{36} < \overline{X} < 800 + 2{,}58\sqrt{36}$
On a donc 99% de chances pour que $784{,}52 < \overline{X} < 815{,}48$

Proportions

∗ Dans la population la fréquence du caractère λ est connue P(λ) = p
∗ Dans un échantillon de taille N , X est le nombre de personnes possédant le caractère λ .
∗ X est la somme de N variables de Bernouilli indépendantes dont la moyenne est p .

∗ **Si N est grand**
Loi des grands nombres : si **N** → ∞ alors **X/N** → **p** .

E(X) = Np , **V(X) = Npq** (avec q = 1 – p)

Posons **f = X / N** proportion du caractère λ dans des échantillons de grande taille.

E(f) = p et **V(f)** $= \frac{pq}{N}$ **(V(f)** $= \frac{pq}{N}\frac{P-N}{P-1}$ si tirage exhaustif)

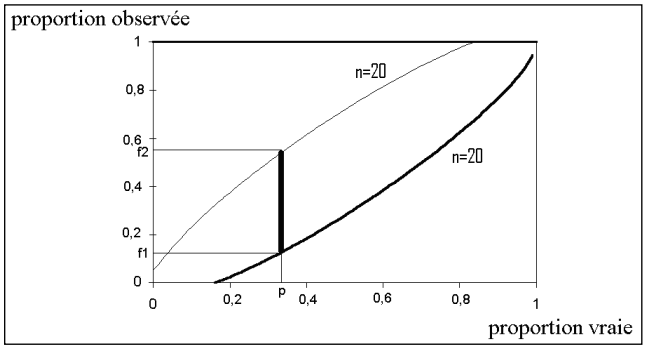
Si N est suffisamment grand on peut avancer que

f obéit à une loi **N (p ,** $\sqrt{\frac{pq}{N}}$**)**

∗ **Si n n'est pas assez grand**
On utilise la loi binomiale **B(N, p)** .

On connaît la proportion vraie (p). Quel est selon la taille de l'échantillon l'intervalle de situation de la proportion observée (f) au risque de r% ?

il existe des tables donnant les intervalles de pari d'une proportion à **95%** (ou à **99%**).
Par exemple, pour un pari à 95%, on peut s'appuyer, entre autres, sur une famille de courbes paramétrées par le nombre d'éléments N de l'échantillon.
Pour chaque valeur de N on a 2 courbes qui ressemblent à celles qu'on voit sur le graphique ci-dessous (Celles – là correspondent à N = 20).



Si la proportion vraie est **p** , on trace la droite **x = p** , elle intercepte le couple de courbes **N = 20** en deux points d'ordonnée **f1** et **f2**. Et on peut dire qu'au risque de **5%** la fréquence **f** qu'on va mesurer dans une échantillon de **20** éléments sera telle que **f1 < f < f2** .
Les courbes de paramètre **N = 50** seront situées entre les courbes **N = 20** ce qui fait, qu'au même risque l'intervalle de confiance rétrécit quand la taille de l'échantillon augmente.

On connaît la proportion vraie (p). Quelle est la taille N de l'échantillon qui permet d'obtenir pour la fréquence observée (f) une précision de e% au risque de r% ?

p = 1/6 (donc q = 5/6)
La fréquence **f** doit être connue au 50e prés (**e% = 2%**) avec une probabilité de **99%** (risque **r% = 1%**) .
Quelle doit être la taille **N** de l'échantillon choisi pour déterminer **f** ?

f obéit à une loi **N (p ,** $\sqrt{\frac{pq}{N}}$**)**

$\frac{f-p}{\sqrt{\frac{pq}{N}}}$ obéit à une loi **N(0,1)**

On a donc **– 2,58 <** $\frac{f-p}{\sqrt{\frac{pq}{N}}}$ **< +2,58** avec une probabilité minimale de **0,99 (**d'après les tables**)**

Ce qu'on peut aussi écrire **P (p – 2,58** $\sqrt{\frac{pq}{N}}$ **< f < p + 2,58** $\sqrt{\frac{pq}{N}}$**) > 0,99**

Il faut donc **2,58** $\sqrt{\frac{pq}{N}}$ **< e%** **ou** **2,58** $\sqrt{\frac{pq}{N}}$ **< 2/100** avec **pq = 5/36**

ce qui donne N > 2311

Estimateurs

On observe N fois la variable **X** d'une population.
On trouve les valeurs X_1 , X_2 , ..., X_N
Nous cherchons à connaître soit la moyenne (**Y =** $\overline{X}$) soit l'écart type (**Y = σ**) de **X** dans la population totale.
Pour cela nous utilisons une valeur, **Tn** , calculée à partir de X_1 , X_2 , ..., X_N

On dit que **Tn** est un **estimateur** de **Y** si **Tn** converge en moyenne quadratique vers **Y**
c'est-à-dire si lorsque **N** $\rightarrow \infty$
E(Tn) $\rightarrow$ **Y**
V(Tn) $\rightarrow$ **0**
Si **E(Tn) =Y** l'estimateur est dit « **sans biais** »

■ **On appelle biais de Tn pour θ la valeur** $b_\theta(Tn) = E[Tn] - \theta$.
L'estimateur Tn sera dit sans biais si E [Tn] = θ. Sinon on dit qu'il est biaisé .
■ **Un estimateur Tn est dit asymptotiquement sans biais si E [Tn] tend vers θ lorsque n** $\rightarrow \infty$ **.**
■ **Un estimateur Tn est consistant (convergeant) si Tn converge en probabilité vers θ lorsque n**$\rightarrow\infty$
■ **Si Tn est asymptotiquement sans biais et de variance tendant vers 0 lorsque n** $\rightarrow \infty$ **, alors Tn est consistant.**
■ **L'erreur quadratique moyenne est le moment d'ordre 2 de l'erreur d'estimation :** $E[(Tn-\theta)^2]$
■ **Si Tn est de carré intégrable, l'erreur quadratique moyenne se décompose en un terme de biais et un terme de variance** $E[(Tn-\theta)^2] = (E[Tn]-\theta)^2 + Var(Tn)$.

∗ cas d'un tirage avec remise : **f** est un estimateur sans biais de la proportion **p** réelle dans la population.
∗ Mode de tirage quelconque : E**(**$\overline{X}$**) = m** (moyenne réelle de X) . $\overline{X}$ est un estimateur sans biais de **m.**
∗ Tirage sans remise : Par contre **V(X)** n'est pas un estimateur sans biais de σ^2 .
∗ Tirage sans remise : $S^2 = \frac{N}{N-1}V(X)$ est un estimateur sans biais de σ^2**.**

$S = \sqrt{\frac{\sum(X_i-\overline{X})^2}{N-1}}$ **.** **S** et σ équivalents si N est grand

Problème : Quelle est la précision de l'estimation ?

Estimation.

On procède à une estimation lorsque ayant affaire à une population peu ou mal connue, on cherche à évaluer certaines de ses caractéristiques (moyenne, proportion, écart type) à travers les caractéristiques d'un échantillon de taille N.

Estimation de la moyenne

P population
Si P suit une loi normale → $\overline{X}$ moyenne de l'échantillon suit une loi normale
Si P inconnue et N effectif de l'échantillon **> 30** → $\overline{X}$ moyenne de l'échantillon suit une loi normale
Si P inconnue et N effectif de l'échantillon **< 30** → loi de Student avec N – 1 degrés de libertés.

∗ Effectif de l'échantillon N > 30 , σ (population) connu

$$\overline{X} - 1{,}96\,\frac{\sigma}{\sqrt{N}} < m < \overline{X} + 1{,}96\,\frac{\sigma}{\sqrt{N}}$$

De **-1,96 <** $\frac{\overline{X}-m}{\frac{\sigma}{\sqrt{N}}}$ **<+1 ,96** avec une probabilité de 95% on tire
ce qui donne un encadrement de **m** au risque de 5%

∗ Effectif de l'échantillon N > 30 , σ (population) inconnu

L'écart type σ n'étant pas connu, on en fait une estimation : $S = \sqrt{\frac{\sum(X_i-\overline{X})^2}{N-1}}$
(ou si on connaît la variance dans l'échantillon : **S =** $\sqrt{\frac{N}{N-1}V(X)}$ **)**

A partir de là , on procède comme précédemment :

$$\overline{X} - 1{,}96\,\frac{S}{\sqrt{N}} < m < \overline{X} + 1{,}96\,\frac{S}{\sqrt{N}}$$

ce qui donne un encadrement de **m** au risque de 5% **.**
On remplace 1,96 par 2,33 pour un encadrement au risque de 1%.

∗ Effectif de l'échantillon N < 30 , Test de Student

Supposons qu'on veuille un encadrement de la moyenne au risque de **1%.**
∗ On cherche dans la table de Student : la valeur de **t** qui a **1%** de chance d'être dépassé avec **N – 1** degrés de liberté est **T**.
∗ On calcule la moyenne $\overline{X}$ de l'échantillon.
∗ On estime l'écart type **S** de la population $S = \sqrt{\frac{\sum(X_i-\overline{X})^2}{N-1}}$
∗ Et on écrit l'encadrement au risque de 1% :

$$\overline{X} - T\,\frac{S}{\sqrt{N}} < m < \overline{X} + T\,\frac{S}{\sqrt{N}}$$

Estimation d'une proportion

∗ Dans la population la fréquence **p** du caractère λ est inconnue
∗ Dans un échantillon de taille **N** , **X** est le nombre de personnes possédant le caractère λ
∗ **f = X / N** est la fréquence observée dans l'échantillon
∗ Il faut donner une estimation de **p.**

X suit une loi binomiale **B (N, p)**
f suit une loi parente de la loi binomiale
P (X = x) = P(f = x/N) donc
E(f) = p et **V(f) =** $\frac{p(1-p)}{N}$ avec **p** inconnu

∗ **p** est estimé par **f** (fréquence dans l'échantillon)
∗ **V(p)** est estimé soit par $\frac{f(1-f)}{N-1}$ (N grand)

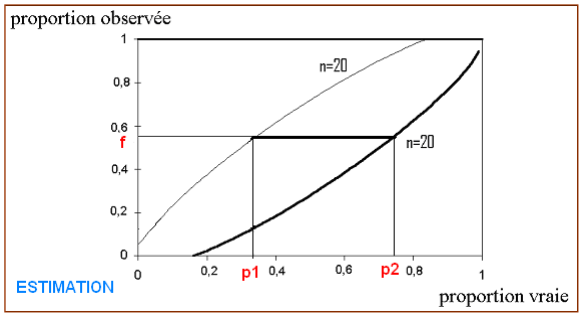
soit par $\frac{1}{4(N-1)}$ valeur maximale de la variance obtenue pour f = 1 – f = ½ (N petit)

∗ Si **N > 100** la loi de **p** est approchée par une loi normale **N (f ,** $\sqrt{\frac{f(1-f)}{N-1}}$**)**

Pratiquement

∗ Si N est petit

on utilise à l'envers l'abaque qu'on utilisait pour encadrer la fréquence observée connaissant la proportion réelle.



On choisit la famille de courbes correspondant au risque **r%** choisi (5% ou 1% en principe)
Dans cette famille on choisit le couple de courbes paramétrées par **n** le plus proche possible de **N** (ici n = 20)
On trace la droite y = f qui coupe le couple de courbes de paramètre **n** en des points d'abscisse **p1** et **p2** .
On estime qu'au risque de **r%** la proportion vraie, **p**, vérifie **p1 < p < p2**

∗ Si N est grand

On suppose que **p** est distribué selon une loi normale **N (f ,** $\sqrt{\frac{f(1-f)}{N-1}}$**)**

On calcule **s =** $\sqrt{\frac{f(1-f)}{N-1}}$ estimation de l'écart type

Et on écrit par exemple qu'au risque de 5%
f – 1,96 S < p < f + 1,96S

Estimation d'un écart type

Ponctuellement **S** est estimé par $S = \sqrt{\frac{\sum(X_i-\overline{X})^2}{N-1}}$

σ est estimé par **σ =** $\sqrt{\frac{\sum(X_i-m)^2}{N}}$ (m de la population connue) ou $\sqrt{\frac{\sum(X_i-\overline{X})^2}{N}}$ (m inconnue)

L'estimateur suit une loi **A = (n – 1)** $\frac{s^2}{\sigma^2}$ du χ^2 à **N-1** degrés de libertés

1) On calcule **a = (n – 1)** $\frac{s^2}{\sigma^2}$

2) On cherche pour **N – 1** degrés de liberté la probabilité **P(A > a)** (table du χ^2)

3) si on a 0,025 < P(A > a) < 0,975 on dira qu'on a un intervalle de confiance de **5%**

4) Si χ^2 **(0,025) = a1** et χ^2 **(0,975) = a2** on peut dire que **a1 < A < a2** avec une probabilité de **5%**

5) de **a1 <** $\frac{s^2}{\sigma^2}$ on tire $\sigma < \sqrt{\frac{(n-1)s^2}{a1}}$

6) de **(n – 1)** $\frac{s^2}{\sigma^2}$ **< a2** on tire $\sigma > \sqrt{\frac{(n-1)s^2}{a2}}$

7) et en fin de compte on a l'encadrement de **σ** suivant au risque de 5% :

$$\sqrt{\frac{(n-1)s^2}{a2}} < \sigma < \sqrt{\frac{(n-1)s^2}{a1}}$$

Complément

Emprunt à WIKIPEDIA

Lorsqu'il s'agit d'estimer la dispersion autour de la moyenne d'un caractère statistique dans une population de grande taille à partir d'un échantillon de taille n, on utilise pour l'écart type la valeur suivante

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

.

On peut remarquer que

$$s = \sigma\sqrt{\frac{n}{n-1}}$$

Pourquoi n-1 ?

La question que l'on se pose généralement est « Pourquoi n - 1 ? ». La raison pour laquelle on divise par n - 1 au lieu de n est un bel exemple de l'interaction permanente entre les statistiques et les probabilités.

Le sondage de n individus correspond à une série de n variables aléatoires xi indépendantes d'espérance E(X) et de variance V(X).

La moyenne $\overline{x}$ de l'échantillon est une variable aléatoire d'espérance E(X) et de variance

$$\frac{1}{n} \cdot V(X)$$

(la moyenne de n variables aléatoires fluctue moins qu'une seule variable aléatoire).

La variance v de l'échantillon est une variable aléatoire dont on veut calculer l'espérance.

$$v = \left(\frac{1}{n}\sum x_i^2\right) - \overline{x}^2$$

.

x_i^2 est une variable aléatoire d'espérance

$$E(x_i^2) = E(x_i)^2 + V(x_i)$$

donc égale à E(X)2 + V(X).

$$\frac{1}{n}\sum x_i^2$$

est une variable aléatoire d'espérance $E(X)^2 + V(X)$.

$\overline{x}^2$ est une variable aléatoire d'espérance

$$E(\overline{x})^2 + V(\overline{x}) = E(X)^2 + \frac{1}{n}V(X)$$

.

Donc

$$E(v) = E(X)^2 + V(X) - E(X)^2 - \frac{1}{n}V(X) = \frac{n-1}{n}V(X)$$

.

La variance v de l'échantillon fluctue donc autour de

$$\frac{n-1}{n}V(X)$$

et non autour de V(X) comme on aurait pu s'y attendre.

Pour obtenir une estimation de V(X), il est donc nécessaire de prendre

$$\frac{n}{n-1}v$$

.

On pourrait dire que v est un estimateur biaisé.

Et pour obtenir une estimation de l'écart type σ(X), il est nécessaire de prendre

$$\sigma\sqrt{\frac{n}{n-1}}$$

.

DECISION STATISTIQUE

Comparaison d'une moyenne à une norme m_0

∗ Une hypothèse de comparaison de la moyenne **m** de la population à une norme $\mathbf{m_0}$ est formulée.
Cette hypothèse peut revêtir plusieurs formes
H0 : peut – on dire que $\mathbf{m = m_0}$ égalité
H1 : peut on dire que $\mathbf{m > m_0}$ ou $\mathbf{m < m_0}$ majoration minoration
L'hypothèse est forcément formulée avec un certain risque (en général 5% ou 1%) qu'il nous appartient de préciser.

∗ On peut à l'inverse avoir le certitude que $\mathbf{m = m_0}$ (ou $\mathbf{m > m_0}$ ou $\mathbf{m < m_0}$) et se demander si la procédure de construction de l'échantillon ou les mesures nécessaires à sa constitution sont valables (là aussi la réponse à cette question est formulée au risque de r%) .

Il y a plusieurs cas de figures

∗ Ecart type σ de la population connu, la population suit une loi normale ou n > 30

alors $\overline{X}$ de l'échantillon suit une loi normale $\mathbf{N(\ m\ ,\ \frac{\sigma}{\sqrt{N}}\ )}$

Soit on suppose que $\mathbf{m = m_0}$ et on vérifie que $\overline{X}$ est dans l'intervalle de confiance au risque donné
Soit on formule l'hypothèse que $\mathbf{m > m_0}$ ou $\mathbf{m < m_0}$ et **m** étant encadré en fonction de $\overline{X}$, on regarde où se trouve $\mathbf{m_0}$ par rapport aux bornes de l'encadrement.

∗ Ecart type σ de la population connu, n < 30

Impossible de conclure sauf si l'on connaît la loi de la population

∗ Ecart type σ de la population inconnu, la population suit une loi normale ou n>30

On estime σ par $S = \sqrt{\frac{\sum(X_i - \overline{X})^2}{N-1}}$

La moyenne suit une loi de Student à N-1 degrés de libertés**.**
Si N > 30 la loi de Student peut être approchée par une loi normale $\mathbf{N(\ m\ ,\ \frac{\sigma}{\sqrt{N}}\ )}$
Voir à la fin de ce chapitre.

∗ Ecart type σ de la population inconnu, n < 30

Impossible de conclure

Exemples :

∗ On mesure N fois une longueur l = l_0 (l_0 connue) . On connaît l'écart type sur la mesure σ .
Au risque de 5%, les mesures sont elles réalisées convenablement ?
C'est le cas si $\mathbf{l_0 - 1{,}96\frac{\sigma}{\sqrt{N}} < \overline{X} < l_0 + 1{,}96\frac{\sigma}{\sqrt{N}}}$

∗ On mesure 100 fois une valeur X . On fixe une norme $\mathbf{m_0}$ pour $\overline{X}$. Peut on dire qu'en moyenne $\overline{X}$ne dépasse pas $\mathbf{m_0}$ avec un risque **1% ?**
On ne connaît pas **σ** de la population mais on l'estime grâce à notre échantillon $\boldsymbol{\sigma = s}$.
Pour N = 100 on peut approcher la loi de Student par une loi normale.
On évalue la moyenne $\overline{X}$de l'échantillon.
L'hypothèse est acceptable si $\mathbf{\overline{X} < m_0 + 2{,}33\frac{\sigma}{\sqrt{N}}}$

Comparaison d'une fréquence à une norme p_0

Le problème est le même que pour une moyenne
Si l'on s'intéresse à la fréquence p de la modalité λ d'un caractère.
Si **X** est le nombre d'individus pour lesquels le caractère est **λ** dans un échantillon de grandeur **N .**
X suit une loi binomiale B(N, p_0) que l'on approche par une loi normale dés que **N** est assez grand.

Pour ce qui est de **f = X / N** à la limite sa loi normale est $\mathbf{N(p0\ ,\ \sqrt{\frac{p_0(1-p_0)}{N}}}$

Exemple
N = 100 et **X = 12** (**f** constatée **= 0,12**).
Peut – on admettre que **p = 1/6** au risque de **5%** (comparaison à p_0 **= 1/6 = 0,17**).
Si l'on admet que la population suit une loi B(100 ; 0,17) on peut l'approcher par la loi normale
N(0,17 ; 0,0395) et au risque de **5%** on devrait avoir
0,17 – 1,96(0,0395) < f < 0,17 + 1,96(0,0395)
soit 0,093 < f < 0,24 et comme **f = 0,12** l'hypothèse est admissible.

Comparaison de deux échantillons

Il s'agit de savoir si avec un risque de r% deux échantillons ont été prélevés dans la même population.

Rappel :
Soit **n** variables aléatoires X_i indépendantes chacune suivant une loi normale d'espérance mathématique m_i et d'écart type σ_i
ΣX_i suit une loi $N(\Sigma m_i ; \sqrt{\sum \sigma_i^2})$

fréquences

1er échantillon effectif **N1** , fréquence mesurée **f1**
2e échantillon effectif **N2** , fréquence mesurée **f2**
Si p_0 est la proportion dans la population **f** d'un l'échantillon issu de la population suit une loi normale de type $N(p_0 , \sqrt{\frac{p_0(1-p_0)}{N}})$

Soit **d = |f1 – f2|** la différence entre les deux fréquences

On estime l'écart type de **d** par $\sigma_d = \sqrt{\frac{f1(1-f1)}{N1} + \frac{f2(1-f2)}{N2}}$

Si **N1 > 30 et N2 > 30** on fait l'approximation normale. **d** doit suivre une loi $N(0 , \sigma_d)$

Donc si les 2 échantillons proviennent de la même population, au risque de r% on doit avoir

$0 - \Delta_r \sigma_d < d < 0 + \Delta_r \sigma_d$ (avec par exemple Δ_r = **1,96** pour **r% = 5%)**

Moyennes

1er échantillon effectif **N1** , moyenne mesurée **m1** , écart type mesuré **s1**
2e échantillon effectif **N2** , moyenne mesurée **m2** , écart type mesuré **s2**
Si **m** est la moyenne et **σ** l'écart type dans la population, la moyenne $\overline{X}$ des échantillons tirés de cette population suit une loi normale de type $N(m, \sqrt{\frac{\sigma^2}{N}})$

Soit **d = |m1 – m2|** la différence entre les deux moyennes

Si **N1 < 30 et N2 < 30** et que les populations d'origines sont distribuées normalement
On applique une **loi de Student** à **n1 + n2 – 2** degrés de liberté

Si **N1 > 30 et N2 > 30** que les populations d'origine soient normales ou pas, on fait l'approximation normale.

$S_d = \sqrt{\frac{{s_1}^2}{N1} + \frac{{s_2}^2}{N2}}$

d doit suivre une loi $N(0 , S_d)$

Donc si les 2 échantillons proviennent de la même population, au risque de r% on doit avoir
$0 - \Delta_r S_d < d < 0 + \Delta_r S_d$ (avec par exemple Δ_r = **2,58** pour **r% = 1%)**

Comparaison de deux distributions

Test du $\chi 2$

On utilise ce test pour juger
∗ De l'adéquation d'une population à une distribution type (exemple loi de poisson)
∗ De l'homogénéité de 2 populations soupçonnées de suivre une même loi
∗ De l'indépendance de deux populations

Méthode

On répartit les valeurs de l'échantillon (de taille N) dans **k** classes distinctes et on calcule les effectifs de ces classes. Si l'on regroupe certaines classes pour les doter d'un effectif plus important, **k** diminue en conséquence.
Appelons o_i (i=1,...,k) les effectifs observés et $\mathbf{T_i}$ les effectifs théoriques.
On calcule

$$\chi^2 = \sum \frac{(O_i - T_i)^2}{T_i}$$

La statistique χ^2 donne une mesure de l'écart existant entre les effectifs théoriques attendus et ceux observés dans l'échantillon. En effet, plus χ^2 sera grand, plus le désaccord sera important. La coïncidence sera parfaite si χ^2**=0.**

Le degré de liberté (d) de la variable soumise au test ($\mathbf{o_i}$) est obtenu en soustrayant à **k** le nombre de relations entre les **k** valeurs qui ont été utilisées dans le paramétrage de la loi de référence.
Par exemple si on a une relation de type $\mathbf{\Sigma\ o_i = n}$ (loi B(n,p)) le degré de liberté devient **k – 1**
Si, de plus on a eu besoin de calculer **la moyenne m** des $\mathbf{o_i}$ pour tester l'adéquation à la loi B(n,p) (p déduit de m) ou P(m) (m paramètre de la loi de Poisson) le degré de liberté deviendra **k – 2.**

La table donne, pour **d** degrés de libertés, une fonction de répartition de χ^2 : la probabilité pour que χ^2 soit plus grand qu'une valeur donnée **q**.
En situant χ^2 dans l'échelle des valeurs de **q** on sait que χ^2 a entre **x%** et **y%** de chances d'être dépassé. Plus la probabilité de χ^2 d'être dépassé est grande, plus l'adéquation de la loi à la série est judicieuse

Pour 5 degrés de libertés	
P(X²> q)	**q**
0.9	**1,61**
0.8	**2,34**
0.7	**3**
0.5	**4,35**
0.3	**6,06**
0.2	**7,29**
0.1	**9,23**
0.05	**11,07**
0.02	**13,39**
0.01	**15,08**

Exemple

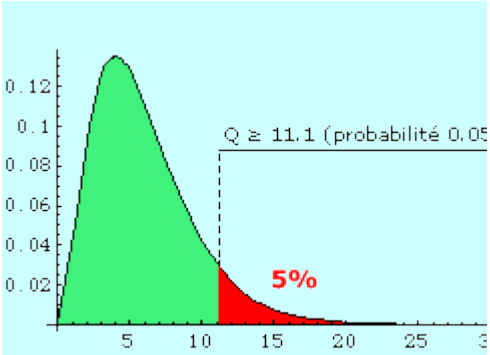
On a lancé un dé 90 fois et on a obtenu les issues 1 à 6 (k=6) avec les effectifs suivants: 12, 16, 20, 11, 13, 18. Si le dé n'est pas pipé (notre hypothèse), on attend comme effectifs moyens théoriques 15 pour toutes les issues. On pose χ^2= Q

$$Q = \frac{(12-15)^2}{15} + \frac{(16-15)^2}{15} + \frac{(20-15)^2}{15} + \frac{(11-15)^2}{15} + \frac{(13-15)^2}{15} + \frac{(18-15)^2}{15} = \frac{64}{15} = 4.2($$

Pour k-1=5 degrés de liberté on trouve dans la table Q entre les valeurs

0.7	**3**
0.5	**4,35**

Ce qui signifie que la probabilité pour Q d'être dépassé est un peu supérieure à 50% . L'adéquation de la loi à la série n'est pas fameuse.



Fonction de répartition de la loi du χ^2 pour 5 degrés de liberté.

Exemples :

1
Les **Oi** sont les effectifs des classes pour lesquelles **X = Xi** .
On a une série de **7** valeurs **Oi** avec pour seule relation $\mathbf{\Sigma\ O_i = 100.}$
Donc **d = 7 – 1 =** 6 degrés de liberté.
On nous demande si on peut ajuster cette série par une loi **B(100 ; 0,04).**
On calcule les 7 effectifs théoriques correspondants $\mathbf{T_i}$.
Puis χ^2 . On trouve χ^2 **= 0,45**
Dans la table pour **d = 6** on lit **P(χ^2 > 2,2) = 0,9**
Donc à fortiori **P(χ^2 > 0,45) > 0,9**
On en déduit qu'on peut sans problème ajuster la série par la loi **B(100 ; 0,04).**

2

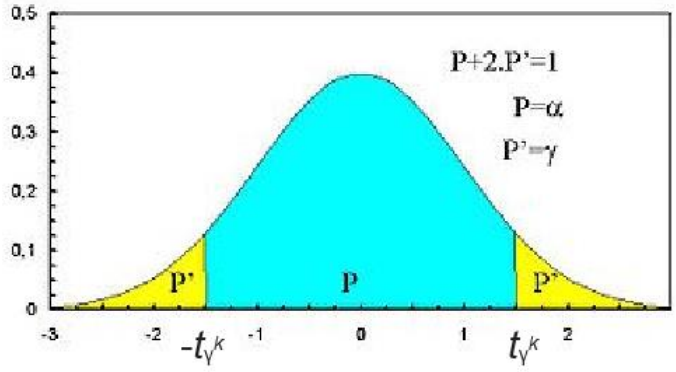
Pour la même série, on calcule la moyenne **m** de l'échantillon.
m = 3,9 pour un effectif de **100.**
On nous demande si l'on peut ajuster la série par la loi **B(100 ; 0,39)**
Cette fois on a 2 relations entre les paramètres de la loi et les données:
$\Sigma o_i = 100$
et **$0,39 = (\Sigma o_i X_i) / 100$**
Donc il faudra regarder dans la table de **d =7 – 2 =** 5 degrés de liberté.
Où l'on trouvera **$P(\chi^2 > 1,6) = 0,9$**
Et comme nous calculerons un χ^2de **0,4** nous en déduirons que là aussi l'ajustement est acceptable.

3

Puis pour la même série on peut tenter l'ajustement par une loi normale.
N(m , σ) .
Pour cela il faut calculer **σ** ce qui nous donne une 3e relation aux côtés de
$\Sigma o_i = 100$ et **$0,39 = (\Sigma o_i X_i) / 100$**
On aura donc **d = 7 – 3 =** 4 degrés de libertés.
Dans la table on trouvera **$P(\chi^2 > 1,06) = 0,9$**
On voit que plus l'ajustement est ambitieux (accroissement du nombre de paramètres exigés par la loi) plus l'exigence en χ^2 est sévère (plus le χ^2doit être faible)

Utilisation loi de Student à k degrés de liberté.

Si Z suit une loi normale centrée réduite et U distribuée selon la loi du KHI2 à k degrés de liberté T = $\frac{Z}{\frac{U}{k}}$ suit une loi de Student à k degrés de liberté.



Pour k degrés de liberté, les tables de Student donnent $\mathcal{T} = t_r^k$ qui est la valeur de t pour laquelle P(t > $\mathcal{T}$) = r (r étant le risque de l'évaluation)

Si x1,…xn suivent une loi normale d'espérance e (à déterminer) et de variance σ^2 (inconnue) , au niveau de confiance c.

Autrement dit on veut connaître l'intervalle de confiance de e au risque de 1-c.

Au risque de 1– c, **e** appartient à l'intervalle $\left[\bar{x} - t_{\frac{1-c}{2}}^{n-1}\sqrt{\frac{S}{n}} \;,\quad \bar{x} + t_{\frac{1-c}{2}}^{n-1}\sqrt{\frac{S}{n}}\right]$

Avec x =moyenne des Xi et S estimateur de σ = $\frac{\sum(Xi-\bar{X})^2}{n-1}$