CHAPITRE V LE JUGEMENT SUR ECHANTILLON ESTIMATION ET TEST D'HYPOTHESE

A la recherche de la pièce parfaite…

Nous aimerions qu'une pièce parfaite ait les piles et les faces également distribuées. Le tableau ci-dessous donne les résultats de 500 jets successifs d'une pièce de monnaie, s'agit-il des jets d'une pièce non truquée ?

P	P	F	P	F	P	F	F	F	P
F	P	P	F	F	P	P	P	F	P
P	P	F	P	F	P	F	P	P	F
F	F	F	P	F	F	F	P	F	F
F	F	P	P	P	F	F	P	P	F
P	F	F	P	F	F	F	F	F	P
F	P	P	P	P	F	F	F	P	P
F	F	P	F	F	F	F	F	F	F
F	P	P	F	F	P	F	P	P	P
P	F	P	P	F	P	F	F	P	F
P	P	P	P	F	P	F	F	P	P
P	P	F	P	F	P	F	P	P	P
P	P	F	F	F	F	F	F	F	P
P	P	F	F	F	F	F	F	P	F
F	P	P	P	F	P	P	F	F	P
P	F	F	F	P	P	P	F	F	F
P	F	F	P	P	P	P	P	P	P
P	P	F	F	F	P	P	F	F	P
P	P	P	F	F	P	P	P	P	P
P	F	F	P	P	P	P	F	P	P
P	P	F	F	P	F	P	P	P	F
P	P	F	P	P	F	P	F	P	F
F	F	F	P	P	F	P	P	P	F
P	F	P	P	F	P	F	F	P	P
P	F	P	P	F	F	F	P	P	F
P	F	F	P	F	P	F	F	F	F
F	P	F	P	F	F	P	F	P	F
P	P	F	P	F	F	F	F	F	P
F	P	P	F	F	F	P	P	F	F
F	P	P	P	F	P	F	F	P	P
P	P	F	P	F	F	P	F	F	F
P	P	P	P	P	P	F	F	P	F
F	P	P	F	F	F	P	P	F	F
F	P	P	P	P	P	F	P	F	F
P	P	P	F	P	P	P	F	F	P
F	F	P	P	F	F	F	P	F	P
F	P	F	P	F	F	P	F	P	P
F	P	P	P	F	P	F	P	F	P
F	P	F	P	F	F	F	F	F	F
P	F	F	F	F	F	P	F	P	P
F	F	P	F	P	F	F	P	F	P
F	P	P	F	P	F	F	P	F	F
P	P	F	P	P	P	F	F	F	F
F	P	F	F	P	F	F	P	P	F
P	P	P	P	F	F	F	P	P	P
P	F	F	F	F	F	P	P	P	P
F	F	F	P	P	F	F	F	P	P
P	P	F	F	P	F	F	P	F	P
P	P	P	F	P	P	F	F	P	F
P	P	P	F	F	F	P	P	F	P

On a enregistré 251 piles et 249 faces, soit une proportion de piles de 0,502. Peut-on dire que la pièce utilisée est parfaite et quelle est la probabilité de se tromper lorsque l'on fait cette hypothèse ? Ce chapitre se propose de répondre à ces questions.

A) GENERALITES

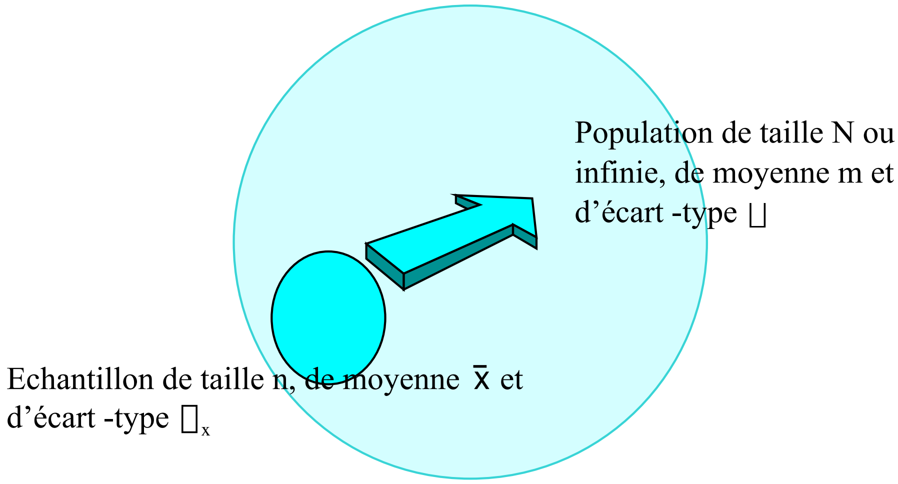
Pour recueillir des informations sur une population, on peut disposer de 2 méthodes :

- La méthode exhaustive ou recensement : on examine chacun des individus de la population selon le caractère étudié (exemple recensement général de la population française)

- La méthode des sondages : on examine une partie de la population appelée échantillon (exemple dans une fabrique d'allumettes, on prélève à la sortie de fabrication des échantillons pour contrôler la qualité du produit)

En général on rejette la première méthode qui est trop onéreuse et même destructrice !!

Le problème de l'estimation est de déduire des informations sur la population à partir des informations données par l'échantillon. Il est clair que lorsqu'on fait un sondage, on ne peut aboutir à une connaissance parfaite des paramètres de la population.
Par exemple, on étudie une variable aléatoire X relative à la population.



On cherche à estimer les paramètres m et σ inconnus à partir des paramètres de l'échantillon avec une précision connue·

La méthode des sondages aléatoires est caractérisée par le fait que l'échantillon est désigné de façon à ce que chaque unité de la population ait une probabilité connue, différente de 0, d'être

retenue. En pratique on affecte à chaque unité de la population la même probabilité, la désignation de celui-ci peut être assimilé au tirage de boules dans une urne.
Les tirages peuvent être exécutes de 2 façons différentes :

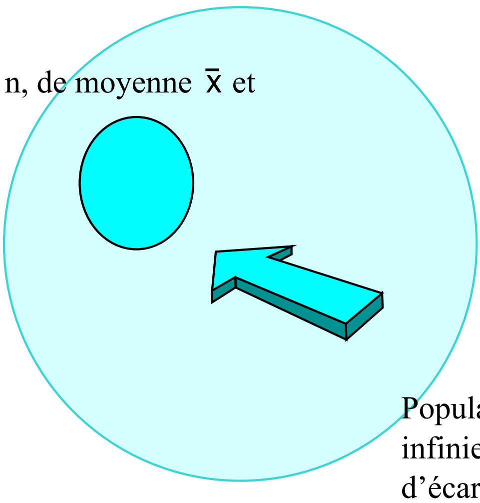
1) Avec remise dans l'urne, les tirages sont indépendants ou non-exhaustifs. Le nombre X d'unités de l'échantillon présentant le caractère déterminé est une VAR binomiale.

2)Sans remise dans l'urne, les tirages ne sont pas indépendants ou exhaustifs. Le nombre X d'unités de l'échantillon présentant le caractère déterminé est une VAR hypergéométrique

Définition : Un échantillon de taille n est la donnée de n observations x_1, x_2, x_3....., x_n de la variable aléatoire X.
On appelle moyenne de l'échantillon le nombre $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ et variance de l'échantillon le nombre $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

On notera que le calcul de la moyenne et de la variance correspond à supposer que les observations sont distribuées de façon uniforme.

Mathématiquement un échantillon est une suite de variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes de même loi que X. On appelle moyenne de l'échantillon la variable aléatoire $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et variance de l'échantillon la variable aléatoire $\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Le processus de décision est fondé sur les propriétés d'échantillonnage, c'est-à-dire, connaissant les caractéristiques de la population , peut-on tirer des informations sur l'échantillon ?



Le fondement de la méthode est la loi des grands nombres et le théorème central limite :

B)____LA LOI DES GRANDS NOMBRES ET LE THEOREME CENTRAL LIMITE

1) LA LOI DES GRANDS NOMBRES (J.BERNOUILLI)

Considérons le tirage d'un échantillon d'effectif n dans une population comprenant des individus ayant un caractère A en proportion p, les individus n'ayant pas le caractère A étant en proportion 1-p. La variable aléatoire X correspondant au nombre d'individus de l'échantillon ayant le caractère A suit une loi B(n, p). La fréquence f_n des individus de l'échantillon ayant le caractère A étant le rapport X/n.

Ainsi, $\mathbf{E}(X/n) = p$ et $\mathbf{V}(X/n) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$

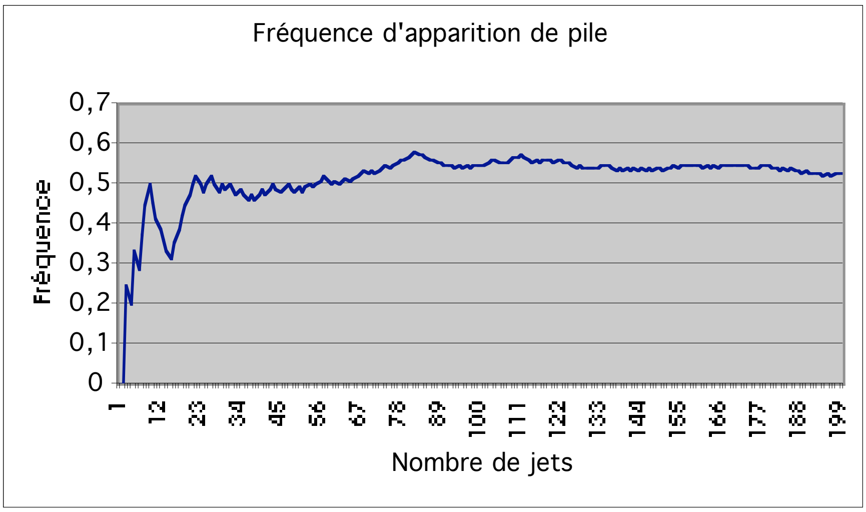
Appliquons l'inégalité de Bienaymé- Tchebychev, on obtient :

Pour tout t dans $\mathbf{R}^{+*}$ $\quad P\left[\ |f_n - p| \geq t\sqrt{\frac{p(1-p)}{n}}\ \right] \leq 1/t^2$

Par suite on peut choisir t assez grand pour que la probabilité considérée soit aussi proche de 0, t étant fixé, on peut choisir l'effectif de l'échantillon n pour que f_n soit aussi voisin de p qu'on le désire. On dit que la fréquence observée converge en probabilité vers p lorsque n tends vers l'infini.

Conséquences : En général, on ne connaît pas la probabilité d'un événement et la loi des grands nombres permet d'estimer cette probabilité à partir de la fréquence d'un échantillon d'effectif suffisamment grand.

Exemple :Fréquence d'apparition de « pile » pour une pièce parfaitement équilibrée jetée n fois, résultat d'une simulation sur ordinateur.



Autre forme de la loi des grands nombres

Théorème : Si $X_1, X_2, \ldots X_n$, est une famille de VAR indépendantes de même moyenne m et d'écart type σ alors la variable aléatoire $\overline{X} = (X_1+X_2+\ldots+X_n)/n$ est une VAR de moyenne m et d'écart -type $\frac{\sigma}{\sqrt{n}}$ et de plus,

$$\text{Pour tout } t \text{ dans } \mathbf{R}^{+*} \qquad P\left[\,|\,\overline{X} - m\,| \leq \frac{\sigma}{\sqrt{n}}t\right] \geq 1 - 1/t^2$$

Il suffit donc de tirer un échantillon de taille n suffisamment grande par rapport à la population pour que la moyenne de la variable soit proche de son espérance mathématique avec une probabilité aussi grande que l'on veut.

2) LE THEOREME CENTRAL LIMITE (ou de LINDEBERG-LIAPOUNOV))

La loi normale est d'application très générale. Elle est en effet engendrée sous des conditions très peu restrictives par l'addition de causes de fluctuations nombreuses et indépendantes. Ceci est dû au théorème central limite .

Théorème : Si X_1, $X_2, \ldots$, X_n, est une famille de VAR indépendantes de moyenne respectivement $m_1, m_2, \ldots, m_n$ et de variance $V_1, V_2, \ldots, V_n$.
Si , lorsque n tends vers l'infini, et pour chaque i, le rapport $\frac{V_i}{\sum_{j=1}^{n} V_j}$ tends vers 0, alors la variable aléatoire moyenne $\overline{X} = (X_1+X_2+\ldots+X_n)/n$ est une VAR normale de moyenne $m = \frac{1}{n}\sum_{i=1}^{n} m_i$ et de variance $\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{n} V_i$.
On dit que, lorsque n tends vers l'infini la VAR $(\overline{X}-m)/\sigma$ converge en loi vers une loi normale N(0,1).

Les conditions du théorème sont en particulier vérifiées lorsque les variables ont même variance ou lorsque les variances sont uniformément bornées.

Un cas particulier courant est le cas d'un échantillon de taille n extrait d'une population de moyenne m et d'écart- type s alors si X représente la moyenne de l'échantillon, on a **E**(X) = m et **V**(X) = σ^2/n.

Les conditions sont en particulier vérifiées lorsque l'on a :

- Une loi binomiale et une grande valeur du paramètre n,
- La moyenne d'un gros échantillon avec remise
- La moyenne d'un gros échantillon sans remise.

Plus précisément:

a) LOI BINOMIALE

Si X suit une loi binomiale B(n , p), on peut considérer que $X = X_1+X_2+....+X_n$ ou les variables X_i sont indépendantes et suivent chacune une loi de Bernoulli B(1 , p).

Ainsi la variable X/n est de moyenne p et de variance $np(1-p)/n^2 = p(1-p)/n$

Si n est assez grand, $\dfrac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}$ qui s'écrit $\dfrac{X-np}{\sqrt{np(1-p)}}$ suit une loi normale réduite centrée, N(0,1).

En pratique on fait l'approximation lorsque np>5 et n(1-p)>5

b) LOI LIMITE D'UNE LOI DE POISSON

Si X suit une loi de Poisson $P(\lambda)$, on peut considérer lorsque $\lambda>18$ que X suit une loi normale $N(\lambda, \sqrt{\lambda})$.

c) LOI DE LA MOYENNE D'UN GROS ECHANTILLON AVEC REMISE.

Un échantillon de taille n est une suite de n variables aléatoires $X_1, X_2,.... , X_n$ indépendantes suivant la même loi que la population parente. La moyenne de l'échantillon est $\overline{X} = (X_1+X_2+....+X_n)/n$.

Lorsque n>30 la moyenne $\overline{X}$ suit une loi normale $N(m, \frac{\sigma}{\sqrt{n}})$.

d) LOI DE LA MOYENNE D'UN GROS ECHANTILLON SANS REMISE.

La tendance de la distribution de la moyenne vers la loi normale subsiste bien que les conditions d'indépendance ne soient plus respectées.

On démontre que l'écart- type de la moyenne est alors $\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$ lorsque N est l'effectif de la population.

Le coefficient correctif $\sqrt{\frac{N-n}{N-1}}$ est le coefficient d'exhaustivité, il tends vers 0 lorsque n augmente(la précision de l'évaluation d'une moyenne augmente au fur et à mesure que l'effectif de l'échantillon se rapproche de celui de la population).

En général, N est grand (condition N>10n) et le coefficient d'exhaustivité est proche de 1.

e) LOI DE LA MOYENNE D'UN PETIT ECHANTILLON.

Lorsque la population suit une loi normale, $\overline{X}$ suit une loi normale N(m, $\frac{\sigma}{\sqrt{n}}$) puisque $\overline{X}$ est une somme de variables normales et indépendantes.

C) L'ESTIMATION

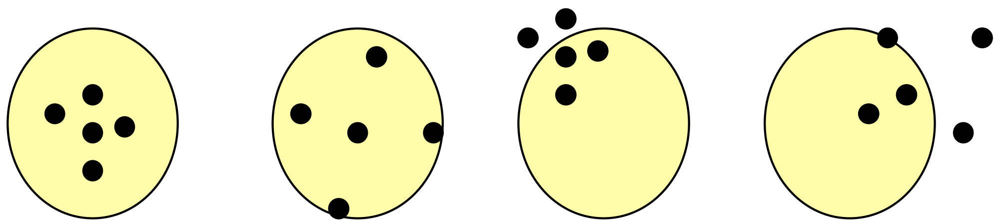
Si $X_1,X_2,....,X_n$ est un échantillon de taille n, d'une population et si $\square$ est un paramètre de cette population(par exemple m, $\square$, p) , on appelle estimateur de $\square$ une fonction $\square(X_1,X_2,....,X_n)$ telle que si n tends vers l'infini, **E**($\square(X_1,X_2,....,X_n)$) tends vers $\square$ et **V**($\square(X_1,X_2,....,X_n)$) tends vers 0.
Un estimateur est dit sans biais lorsque **E**($\square(X_1,X_2,....,X_n)$) = $\square$.

Exemple : $\overline{X}$ est un estimateur sans biais de m. Par contre $\square_X{}^2$ n'est pas un estimateur sans biais de la variance. On montre que **E**($\square_X{}^2$)= $\frac{n\square 1}{n}\square^2$ et par conséquent $S_X^2 = \frac{n}{n-1}\square_X^2$ est un estimateur sans biais de la variance.

Noter qu' un estimateur est une variable aléatoire. Une valeur observée de cette variable est appelée une **estimation** du paramètre.

Lorsque on estime le paramètre $\square$, la variable aléatoire ($\square(X_1,X_2,....,X_n)$ - $\square$ correspond à l'erreur que l'on fait .
Le biais est **E** [($\square(X_1,X_2,....,X_n)$)- $\square$], la racine carrée de **V**[($\square(X_1,X_2,....,X_n)$ - $\square$] est l'erreur type.

La signification concrète du biais et de l'erreur type peut être visualisée de la façon suivante :
Dans une épreuve de tir, la régularité du tireur se manifeste par le fait que les impacts de balle sont groupés (faible erreur type) mais les impacts ne sont pas obligatoirement centrés (notions de biais, erreur systématique)



Absence de biais / Faible erreur type — Absence de biais / Forte erreur type — biais / Faible erreur type — biais / Forte erreur type

On comprend bien sur cet exemple que le statisticien pourra préférer un estimateur avec biais mais de faible erreur type à un estimateur sans biais de forte erreur type.

1) Etude de quelques estimateurs

a) Estimateur de la moyenne d'une population

Il résulte de ce qui précède que $\overline{X}$ est un estimateur sans biais de la moyenne quel que soit le mode de tirage de l'échantillon.

La variance de cet estimateur est égale $\frac{\sigma^2}{n}$ dans le cas de tirages indépendants et à $(\frac{N-n}{N-1})\frac{\sigma^2}{n}$ dans le cas de tirages exhaustifs.

Lorsque l'effectif n de l'échantillon est faible par rapport à celui de la population N, on peut négliger le coefficient d'exhaustivité qui est peu différent de 1(en pratique N>10n).

La loi de $\overline{X}$ n'est en général pas connue, cependant on a les résultats suivants :

- Si σ est connu et $n > 30$ $\overline{X}$ suit une loi normale.
- Si σ est connu et $n \leq 30$ et si la population est normale, $\overline{X}$ suit une loi normale.

b) Estimateur de la variance σ^2 d'une population

Pour estimer la variance de la population, on pense, au premier abord, utiliser comme pour la moyenne la grandeur correspondante, c'est à dire la variance de l'échantillon.

$$\sigma_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Or cet estimateur est biaisé, on montre que $\mathbf{E}(\sigma_X^2) = \frac{n-1}{n}\sigma^2$ et par suite, l'estimateur sans biais de la variance est $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$.

La distorsion provient du fait que, dans ce calcul, les écarts sont mesurés par rapport à la moyenne de l'échantillon et non par rapport à celle de la population.

Dans le cas où l'échantillon est tiré sans remise, on introduit le coefficient d'exhaustivité, l'estimateur de la variance n'est plus S_X^2 mais $\frac{N-n}{N-1}S_X^2$.

Cependant contrairement à ce que l'intuition pourrait suggérer S_X n'est pas un estimateur sans biais de σ.

c) Estimateur d'une proportion

Soit une population comportant deux catégories d'individus, la première correspondant aux cas favorables en proportion p , l'autre en proportion 1-p.

Choisir un échantillon de taille n se ramène à considérer une variable aléatoire X binomiale B(n , p). La variable F= X/n, est un estimateur de p sans biais (c'est la fréquence de cas favorables)

En effet, **E**(X) = np donc **E** (F) = p et **V**(X) = np(1-p) donc **V** (F) = p(1-p)/n.

Dans la cas des tirages exhaustifs X suit une loi hypergéométrique, F est aussi un estimateur sans biais mais **V** (F) = $\frac{N-n}{N-1}$ p(1-p)/n.

Lorsque la taille de l'échantillon est grande on peut supposer que X suit une loi normale et par conséquent F suit une loi normale de moyenne p et de variance p(1-p)/n.
La variance **V** (F) sera estimée par $\frac{F(1-F)}{n-1}$.

2) La recherche d'un estimateur par la méthode du maximum de vraisemblance .

La méthode consiste , étant donné un échantillon , à prendre comme estimation du paramètre θ la valeur de θ qui rend maximale la probabilité de l'événement observé, cette probabilité est appelée la vraisemblance. On montre que parmi les estimateurs non biaisés, l'estimateur obtenu par la méthode du maximum de vraisemblance est celui qui à la plus faible dispersion, c'est à notre point de vue le meilleur estimateur possible.

Exemple 1, cas discret.

On suppose que la population parente suit une loi binomiale B(n ,p) et on cherche un estimateur de p. Si X_1, X_2,.., X_m est un échantillon de taille m et si $k_1, k_2, .., k_m$ est une observation de l'échantillon, la probabilité L= $P[X_1=k_1$ et $X_2=k_2$ et .. et $X_m = k_m]$ est la vraisemblance de l'échantillon.

$$L=\prod_{i=1}^{m} P[X_i = k_i]=\prod_{i=1}^{m} C_n^{k_i} p^{k_i}(1-p)^{n-k_i}$$

L est une fonction de p et la dérivée par rapport à p de Ln(L) est $\sum_{i=1}^{m}\left(\frac{k_i}{p}-\frac{n-k_i}{1-p}\right)$. Cette dérivée s'annule lorsque p=$\frac{\sum_{i=1}^{m} k_i}{mn}$.

Ainsi, L est maximum lorsque p=$\frac{\sum_{i=1}^{m} k_i}{mn}$ et on retrouve le résultat, $\frac{\sum_{i=1}^{m} X_i}{mn}$ est un estimateur de p.

Exemple 2, cas continu.

On suppose que la population parente suit une loi normale N(m, σ) et on cherche un estimateur de m, s étant connu. Si X_1, X_2,.., X_n est un échantillon de taille n et si $t_1, t_2, .., t_n$ est une observation de l'échantillon, la probabilité L= $P[t_1<X_1<t_1+dt_1$ et $t_2<X_2<t_1+dt_2$ et .. et $t_n <X_n <t_n + dt_n]$ est la vraisemblance de l'échantillon.

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(t_i-m)^2}{2\sigma^2}\right)dt_i$$

L est une fonction de m et la dérivée par rapport à m de Ln(L) est $\sum_{i=1}^{n}\frac{(m-t_i)}{\sigma^2}$ .Cette dérivée s'annule lorsque m=$\frac{\sum_{i=1}^{n} t_i}{n}$.

Ainsi, L est maximum lorsque m=$\frac{\sum_{i=1}^{n} t_i}{n}$ et on retrouve le résultat, $\frac{\sum_{i=1}^{n} X_i}{n}$ est un estimateur de m.

Exemple 3

Pour évaluer la population d'un lac, on pêche une première fois mille poissons, que l'on marque et que l'on rejette dans le lac ; puis on procède à une seconde pêche de mille poissons et on constate que k parmi eux sont marqués.

Soit n le nombre de poissons du lac, calculer la probabilité p(n) pour que, lors de la seconde pêche, il y ait exactement k poissons marqués . Donner un estimateur de n par la méthode du maximum de vraisemblance et conclure.

La variable aléatoire X correspondant au nombre de poissons marqués lors de la seconde pêche suit une H(n, 1000, n/1000) et p(n) = $\frac{C_{1000}^{k}C_{n-1000}^{1000-k}}{C_{n}^{1000}}$.

Le rapport $\frac{p(n+1)}{p(n)}=\frac{(n+1-1000)^2}{(n+1)(n+1-2000+k)}$ est supérieur à 1 lorsque (n-999) ^{2} > (n+1)(n-1999+k) soit n<$\frac{-k+1999+(999)^2}{k}$.

Ainsi, p(n) est maximum lorsque n+1=1000000/k et $\frac{10^6}{k}-1$ est un estimateur de n.

3) L'estimation par intervalle.

On veut déterminer l'intervalle de confiance, c'est à dire entre quelles limites se situe la valeur estimée avec une probabilité de se tromper donnée à priori appelé **seuil** α. Pour cela il faut connaître la loi de l'estimateur.

a) Estimation de la moyenne, σ étant connu et n>30.

On sait alors que $\overline{X}$ suit une Loi normale $N(m, \frac{\sigma}{\sqrt{n}})$ et si on se donne un seuil

α = 5% ou 1% alors:

$$P\left[\ |\overline{X}-m| \leq t_\alpha \frac{\sigma}{\sqrt{n}}\right] \geq 1-\alpha$$

$t_{0,05}$= 1,96 et $t_{0,01}$= 2,58

Ainsi, on obtient $P[\overline{X}- t_\alpha \frac{\sigma}{\sqrt{n}} \leq m \leq \overline{X}+ t_\alpha \frac{\sigma}{\sqrt{n}}] \geq 1-\alpha]$ et l'intervalle observé, $[\ \bar{x}- t_\alpha \frac{\sigma}{\sqrt{n}},\ \bar{x}+ t_\alpha \frac{\sigma}{\sqrt{n}}]$ est appelé intervalle de confiance de la moyenne au seuil α.

Il faut noter que l'intervalle de confiance est obtenu à partir d'une estimation on peut donc penser que la vraie valeur de m a une probabilité 1-α de se trouver dans cet intervalle.

b) Estimation de la moyenne, σ étant inconnu et n>30.

Le plus souvent l'écart type σ est inconnu. On l'estime alors en calculant S_X . On obtient l'intervalle de confiance :

$[\ \bar{x}- t_\alpha \frac{S_X}{\sqrt{n}},\ \bar{x}+ t_\alpha \frac{S_X}{\sqrt{n}}]$

c) Estimation de la moyenne, σ étant inconnu et n est petit

Lorsque l'effectif de l'échantillon est faible, la loi normale ne s'applique plus. Si la population parente est normale, on montre que le rapport $\frac{\overline{X}-m}{S_x/\sqrt{n}} = \frac{\overline{X}-m}{\sigma/\sqrt{n}} / \frac{S_x}{\sigma}$ suit une loi de Student à n-1 degrés de liberté, puisque σ est estimé par S_X.

On obtient un intervalle de même type t_α étant remplacé par la valeur s_α extraite de la table .

d)Estimation de la moyenne, σ étant connu et n est petit

On peut faire la même remarque que précédemment, si la population parente est normale, on sait que $\overline{X}$ suit une loi normale. On est donc dans le premier cas.

e)Estimation d'une proportion

Une proportion p est estimée par la fréquence F. F suit une loi parente de la loi Binomiale, F = X/n avec X suit une B(n,p) et ainsi P[F = x] = P[X = nx].

Suivant les valeurs de n on peut supposer que X suit une loi de Poisson ou une loi normale .

Lorsque X est supposée normale, F est aussi normale de moyenne p et de variance $\frac{p(1-p)}{n}$.

On estime $\frac{p(1-p)}{n}$ par $\frac{F(1-F)}{n}$ et si f est la fréquence observée, l'intervalle de confiance au seuil α est donc:

$$[f- t_{\alpha}\sqrt{\frac{f(1-f)}{n}}\ ,\ f+ t_{\alpha}\sqrt{\frac{f(1-f)}{n}}\ ]$$

$t_{0,05}= 1,96$ et $t_{0,01}= 2,58$.

D)__________ DECISION STATISTIQUE, TEST D'HYPOTHESE

Très souvent , on est conduit à confronter une estimation obtenue à partir d'un échantillon à une norme fixée à priori, ou encore à comparer entre eux les résultats de deux échantillons différents. En matière de contrôle de fabrication, on cherchera à déterminer, par exemple, si le diamètre moyen calculé sur un échantillon de pièces mécaniques fabriquées en série est compatible avec la norme spécifiée ou si, au contraire, l'écart observé indique un dérèglement de la machine.

La résolution de ces problèmes de comparaison à partir d'échantillons aléatoires repose sur un mode de raisonnement statistique désigné sous le nom de test d'hypothèse.

1) Principe du test d'hypothèse

Quel que soit le problème posé, les étapes du raisonnement sont les mêmes. Plaçons-nous, par exemple, dans le cas du contrôle par sondage d'une fabrication.
Un échantillon de n pièces a été tiré pour procéder à ce contrôle et on estime qu'une proportion p_0 d'objets non conformes est acceptable. En effet si on veut absolument s'assurer qu'il n'y a aucun objets non conformes, un contrôle exhaustif serait nécessaire.
En général, la proportion observée sur l'échantillon est différente de p_0. Elle pourra, en particulier, lui être supérieure. Cet écart peut avoir deux origines :

- La proportion d'objets non conforme p est bien égale (ou inférieure) à p_0 et la différence observée est due aux seules fluctuations aléatoires, c'est à dire au fait que la mesure est effectuée sur un échantillon.

- La proportion d'objets non conforme est effectivement supérieure à p_0 .

Il s'agit donc de choisir entre ces deux hypothèses et de décider si l'écart observé est **significatif** (au seuil α de probabilité fixé), et la marque d'une différence réelle ou, au contraire, n'est pas significatif et dû seulement au hasard.

On définit les deux hypothèses alternatives H0 et H1 que l'on désire tester :

H0 représente " le pourcentage d'objets non conformes est égal au pourcentage considéré comme acceptable, $p = p_0$ "

H1 représente " le pourcentage d'objets non conformes est supérieur au pourcentage considéré comme acceptable, $p > p_0$ "

On pourrait faire choix d'une autre hypothèse H1 de la forme $p \neq p_0$ mais dans ce cas précis le problème serait mal posé car un pourcentage d'erreur inférieur au pourcentage acceptable constitue une situation favorable.

Le test a pour but de fournir une règle de décision permettant de choisir entre les deux hypothèses.

Si on considère que H0 est exacte, la loi de probabilité de la fréquence mesurée sur l'échantillon est déterminée, c'est suivant le mode de tirage une loi dérivée de la loi binomiale, ou hypergéométrique, loi qui peut être approchée par une loi normale sous certaines conditions.
On se fixe un seuil de probabilité appelé seuil de signification (ou coefficient de risque, le nombre 1- est alors appelé coefficient de sécurité) qui correspond au risque de se tromper que l'on accepte de le courir ; plus précisément a est la probabilité de retenir H1 alors que H0 est vraie:

 = P [choisir H1 / H0 est vraie]

A ce seuil de signification on associe une région critique R et une région d'acceptation (son complémentaire)
La fréquence f observée de l'échantillon appartient soit à la région critique soit à la région d'acceptation :

-Si f appartient à la région critique et si H0 est exacte, il n'y a qu'une probabilité faible a d'observer ce résultat. Il est donc plus probable que H0 soit erronée, on rejette l'hypothèse Ho et on retient l'hypothèse H1.

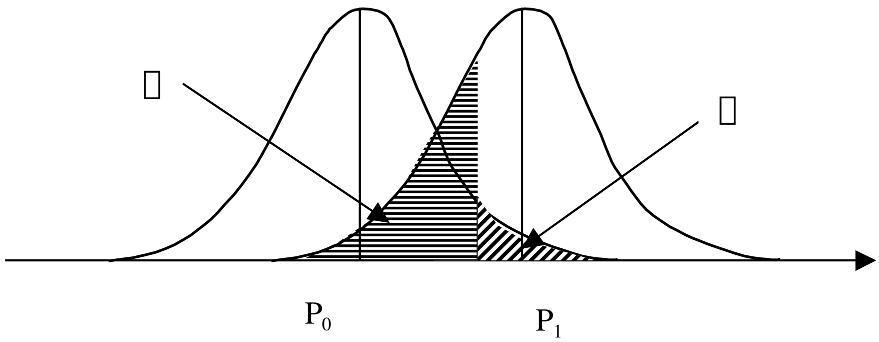
-Si f appartient à la région d'acceptation et si H0 est exacte, la probabilité d'observer ce résultat est élevée 1-. Rien ne s'oppose donc à ce que l'on accepte l'hypothèse H0. Ceci ne prouve pas, cependant que l'hypothèse faite est vraie, mais signifie seulement que les données dont on dispose ne sont pas en contradiction avec celle-ci.

Dans l'exemple introductif, on a trouvé une fréquence observée de 0,502. Si on suppose que la pièce est parfaite (hypothèse H0), la variable aléatoire X nombre de piles suit une loi binomiale B(500, 0,5) et F=X/500.
On peut faire l'approximation normale et ainsi la fréquence F suit une N(0,5, 0, 022). Si on fixe un seuil de 5%, on obtient l'intervalle de confiance P[0,445 < F< 0,544]= 95%.
La valeur observée est dans l'intervalle et il n'y a aucune raison de mettre en doute H0, on peut considérer que la pièce est parfaite.

Remarque :

Dans l'étude d'un test, on a choisi une hypothèse H0 et une hypothèse alternative H1, ainsi qu'un seuil .
Le coefficient de risque , que l'on choisit à priori est la probabilité de se tromper en choisissant l'hypothèse H1, alors que H0 est vraie, a est appelé erreur de première espèce.
On peut aussi se tromper en choisissant l'hypothèse H0, alors que H1 est vraie, la probabilité P[choisir H0/ H1] est notée et appelée erreur de seconde espèce. On montre que lorsque tends vers 0 , tends vers 1.

Contrairement à □, □ n'est pas contrôlé par l'expérimentateur et par conséquent, un choix trop petit de □ augmente les risques de se tromper lorsque l'on accepte H0.

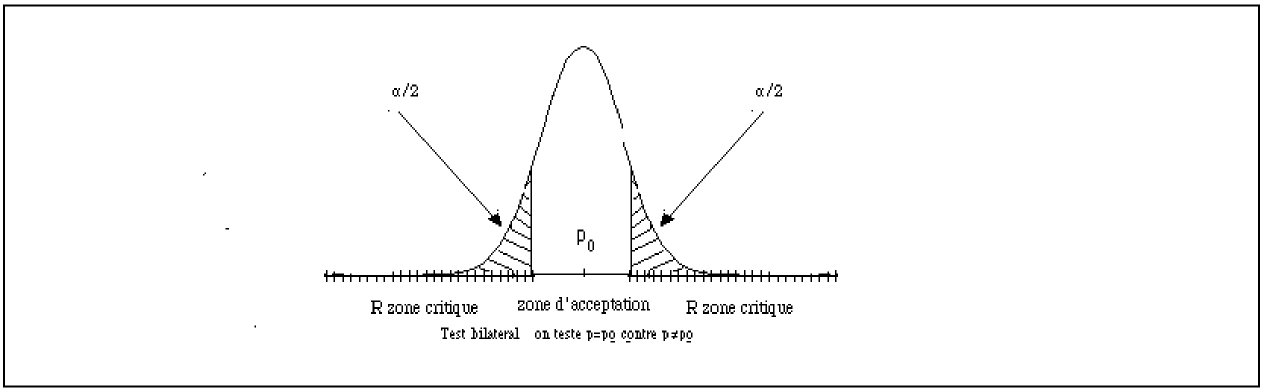


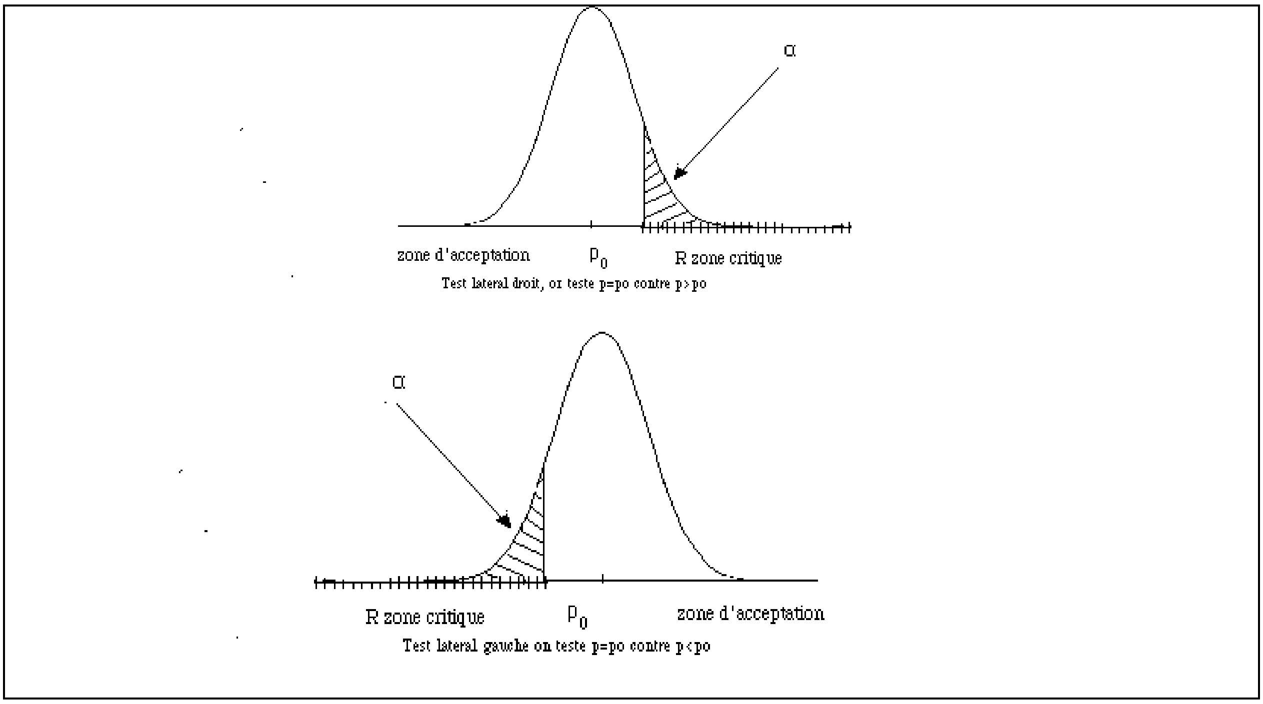
2) Comparaison à un standard

On veut comparer une grandeur estimée p à partir d'un échantillon à une valeur fixée à priori p_0 (standard, norme ,spécification ...)

Ce problème revient au test de deux hypothèses alternatives H0 ($p = p_0$) et H1. L'hypothèse H1 peut prendre des formes différentes suivant la nature de la question posée.

Si H1 est $p \neq p_0$ on dit que le test est bilatéral.
Si H1 est $p < p_0$ on dit que le test est latéral gauche
Si H1 est $p > p_0$ on dit que le test est latéral droit





Exemple 1 On prélève un échantillon de 100 pièces à partir de la production d'une machine. On trouve 12 pièces défectueuses. Peut-on admettre que la proportion de pièces défectueuses sorties de cette machine est de 1 sur 6 au risque 5% ?
La fréquence observée est 0,12, on veut la comparer à 1/6=0,17. En admettant que le nombre de pièces défectueuses suive une B(100, 0,17) hypothèse Ho, l'écart –type de la proportion est $s=\sqrt{\frac{0,17 \times 0,83}{100}}$ =0,037 d'où en faisant l'approximation normale, l'intervalle de confiance est :

$$[0,17 -1,96s \ , 0,17 +1,96s] \text{ soit } [0,097 \ , 0,242]$$

La fréquence observée appartient à l'intervalle, on accepte l'hypothèse.

Exemple2: On mesure une certaine longueur l = 4,5 cm ; on réalise 100 mesures selon une certaine méthode et on fait la moyenne. Quelle règle de décision doit on adopter au seuil de 5% pour décider si ces mesures sont réalisées convenablement, sachant que l'écart -type de la moyenne est s=0,0052 cm ?
Dans le cas des erreurs de mesure , la loi normale s'applique parfaitement. on peut donc supposer que la population est normale de moyenne 4,5 cm, l'écart -type étant connu on a un intervalle de confiance [4,5-1,96s ,4,5+1,96s] soit [4,4898, 4,5102].
On rejettera l'hypothèse que les mesures sont faites correctement si la moyenne observée est en dehors de l'intervalle.

Exemple3: A partir de cent expériences de rupture en charge d'un fil, on a trouvé une moyenne observée de $\bar{x}$= 848 gr et un écart type de σ_X = 71,2.
Peut-on admettre que le poids de rupture ne dépasse pas en moyenne 832 gr avec un risque de 1% ?

Si la population est supposée normale et si on estime s par son estimateur sans biais, la moyenne suit une loi de Student à 99 degrés de liberté, on peut l'approcher par une loi normale.
L'hypothèse Ho est m= 832, l'hypothèse H1 est m> 832. On estime σ^2 par
S_X^2 = 5080,07/0,99=5132,02.
$\overline{X}$ suit une N(832, Sx/10) pour que l'hypothèse soit acceptable il faut que $\bar{x} < 832 + 2{,}33 Sx/10$ soit 848,6 gr l'hypothèse est acceptée.

3) Comparaison de deux échantillons

On se trouve devant deux échantillons dont on ignore le plus souvent s'ils sont tirés de la même population d'origine. On cherche à tester si ces échantillons ont la même caractéristique c . On observe deux valeurs c_1 et c_2 , la différence entre ces deux valeurs peut être due soit aux fluctuations d'échantillonnage soit à la différence des caractéristiques des deux populations d'origine.
On teste donc Ho $c_1=c_2$ contre $c_1 \neq c_2$.
Rappelons quelques propriétés de la différence de deux variables aléatoires:

- Si X et Y sont 2 VAR alors **E**(X - Y) = **E** (X) - **E** (Y)

-Si de plus X et Y sont indépendantes, **V** (X - Y) = **V** (X) + **V** (Y)

- Si de plus X et Y sont normales la différence l'est aussi.

a) Comparaisons de deux fréquences

Soit deux populations A et B composées d'individus ayant un certain caractère en proportion inconnue p et q.
On prélève un échantillon de taille n dans A et de taille m dans B. Sur ces échantillons, on observe les fréquences f et g .
Sur la base de ces observations, on se propose de tester si les proportions peuvent être considérées comme égales.
Les deux hypothèses alternatives sont H_0 " p =q" contre H_1 "p ≠q"

Les fréquences f et g suivent suivant le mode de tirage des échantillons des lois parentes de loi binomiales ou hypergéométriques. Si les effectifs n et m sont suffisamment grands, on peut faire l'approximation normale.
Ainsi f suit une loi normale de moyenne p et de variance p(1-p)/n, g suit une loi normale de moyenne q et de variance q(1-q)/m et par conséquent f -g suit une loi normale de moyenne p -q et de variance $\frac{p(1-p)}{n} + \frac{q(1-q)}{m}$.

Si on suppose que H_0 est vraie alors p =q la différence f- g suit une loi normale de moyenne 0 et de variance p(1-p)$(\frac{1}{m} + \frac{1}{n})$.

On peut donc faire le test et conclure.

Exemple : Pour déterminer le taux d'occupation d'un matériel coûteux, on emploi la méthode des " observations instantanées" : au cours de chaque mois on observe un échantillon d'instants tirés au hasard. Pour chacun de ces instants précis, un contrôleur note si le matériel est utilisé. On a observé un échantillon de 500 instants au mois de janvier et de 400 au mois de février. Les résultats sont les suivants

		Janvier	Février
	Occupation	400	300
	Non occupation	100	100
	Total	500	400

Existe t' il une différence significative entre les taux d'occupation ?
n=500 f= 0,8 m=400 g=0,75 dans l'hypothèse nulle la différence d = f-g suit une loi normale de moyenne 0 et de variance p(1-p)(1/n +1/m) p est estimé par $\frac{nf + mg}{n + m}$ = 0,78 .Au seuil de 5% on obtient un intervalle de confiance pour d de [-2 . 0,028 , +2 . 0,028] la différence observée est à l'intérieur de cet intervalle, elle n'est pas significative. Les observations dont on dispose ne permettent pas d'affirmer que le taux d'occupation du matériel a diminué en Février.

b)Comparaisons de deux moyennes

On considère deux populations P et P'. On prélève un échantillon de taille n dans P et de taille n' dans P'.

Désignons par m et σ la moyenne et l'écart -type de P et m' et σ' la moyenne et l'écart -type de P'.

Les deux hypothèses alternatives sont H_0 "m =m'" contre H_1 "m $\neq$m'"

Si la variable est distribuée dans chaque population suivant une loi normale, les moyennes $\overline{X}$ et $\overline{X'}$ suivent elles-mêmes une loi normale.

Toutefois, si l'hypothèse d'une distribution normale dans les populations ne parait pas justifiée, il suffit que les échantillons soient suffisamment grands.

Sous ces conditions, et en supposant que les tirages peuvent être assimilés à des tirages indépendants :

$\overline{X}$ suit une loi normale de moyenne m et de variance σ^2/n (et respectivement pour $\overline{X}'$).

Par conséquent la différence des moyennes suit une loi normale de moyenne m-m' et de variance $\sigma^2/n + \sigma'^2/n'$.

Sous l'hypothèse H_0 la différence des moyennes suit une loi normale de moyenne 0 et de variance $\sigma^2/n + \sigma'^2/n'$.

On peut donc distinguer deux cas :

- σ et σ' sont connus. le seuil de signification est l'intervalle de confiance correspondant à la valeur de $\sigma^2/n + \sigma'^2/n'$.

-- σ et σ' sont inconnus. on les estimes par S_X et $S_{X'}$.

Exemple:

Dans une grande agglomération, une enquête par sondage a été réalisée sur les dépenses alimentaires mensuelles des ménages. L'échantillon comprenait 327 ménages d'ouvriers et 286 ménages d'employés. On a observé les valeurs suivantes concernant la consommation alimentaire de ces deux catégories sociales :

	Effectif	Moyenne	Ecart -type
Ouvriers	327	612	104
Employés	286	642	118

Peut-on conclure que les ménages d'employés dépensent plus que les ménages d'ouvriers ?
Compte tenu de la taille des échantillons, on peut supposer que les tirages on étés faits avec remise.
Si on fait l'hypothèse qu'il n'y a pas de différence entre les deux consommations, (Hypothèse H_0) la différence entre les moyennes suit une loi normale de moyenne nulle et de variance 81,76 (on estime les variances).
Au seuil de 1% l'intervalle d'acceptation de H_0 est [-23,32 +23,32] la différence observée est -30, elle est située en dehors de l'intervalle et est donc significative.
On peut affirmer, sans beaucoup de chances de se tromper que dans cette agglomération, les employés dépensent en moyenne plus que les ouvriers pour leurs alimentation.

4) Ajustement d'une loi normale à une distribution statistique observée, le test de Khi -deux ou de Pearson

Soit X une variable aléatoire suivant une loi de probabilité théorique P, procéder à N observations de cette variable revient à tirer un échantillon de taille N dans la population théorique infinie correspondant à la loi de probabilité P.
Les observations sont classées suivant k modalités, C_1, C_2,....., C_k qui représentent soit les différentes valeurs possibles, soit les classes de valeurs associées à la variable si celle-ci est continue.
A chacune de ces modalités ou classes correspond une probabilité déterminée par la loi P, p_1, p_2,..., p_k.
L'effectif susceptible d'être observé sur l'échantillon pour chacune de ces classes est X_1, X_2,...X_k.

Ainsi, pour la classe C_i , l'effectif n_i est une variable aléatoire binomiale de paramètre N et p_i, probabilité que la variable appartienne à cette classe. $\mathbf{E}(X_i)=Np_i$ représente l'effectif théorique de la classe C_i, $\mathbf{V}(X_i) = Np_i(1-p_i.)$.
En fait, le k- uplet (X_1, X_2,...X_k) suit une loi multinomiale de paramètre N et p_1, p_2,..., p_k.

Dans ces conditions, et pourvu que la classe C_i soit suffisamment grande pour avoir un effectif théorique d'au moins 4 ou 5 individus (sinon les conditions de convergence ne sont pas remplies), la Variable aléatoire $D = \sum_{i=1}^{k} \frac{(X_i - Np_i)^2}{Np_i}$, une loi de Khi deux à k-1 degrés de liberté.

Cette propriété est tout à fait remarquable, la loi de probabilité de D ne dépend que du nombre de classes et non de la probabilité P ce qui permet de construire un test appelé test de Khi -deux.

LE TEST DE KHI DEUX

Généralement, on ne connaît pas à priori la loi de probabilité suivie par la variable aléatoire X. Suivant la nature du phénomène et après analyse de la distribution observée, on fait choix d'un type de loi dont on estime les paramètres sur la base des observations.(Binomiale, Poisson, Normale....)

Cette loi supposée on choisit k modalités ou classes et on calcule les probabilités p_1, p_2,...,p_k et par suite les effectifs théoriques N,Np_1,Np_2,...,Np_k.

On est alors à même de cal culer une valeur observée d de la variable aléatoire D

$$d=\sum_{i=1}^{k}\frac{(n_i-Np_i)^2}{Np_i}$$

(n_i représente le nombre d'observations d'éléments de la classe i de l'échantillon, N est la taille de l'échantillon)

La variable aléatoire D suit une loi de Khi deux à k-r-1 degrés de liberté , k est le nombre de classes et r le nombre de paramètres estimés.

Le test de Khi -deux repose sur le raisonnement suivant:

On fait l'hypothèse que le phénomène observé suit une loi théorique supposée P, dans ces conditions D est une VAR suivant une loi de Khi -deux à k-r-1 degrés de liberté, d est une valeur observée de D.

- Si il y a une forte probabilité (déterminée par consultation de la table)que D prenne une valeur supérieure à d, l'hypothèse est jugée acceptable.
- Si , par contre, il n'y a qu'une faible probabilité (5% ou 1%) d'obtenir une valeur supérieure à d, il est beaucoup plus probable que cette valeur est due à l'inadéquation de la loi théorique P: on rejette l'hypothèse que le phénomène suive cette loi.

<u>Remarque pratique</u>

- Pour que D suive une loi de khi deux, les conditions de convergence impliquent que les effectifs théoriques des différentes classes ne soient pas trop petits ; en pratique supérieurs à 4.

Par suite, il est parfois nécessaire de regrouper certaines classes notamment aux extrémités de la distribution.

- Habituellement le seuil de probabilité à partir desquels on décide de rejeter l'hypothèse est fixé à 5% ou 1%.

Exemple 1 cas discret.

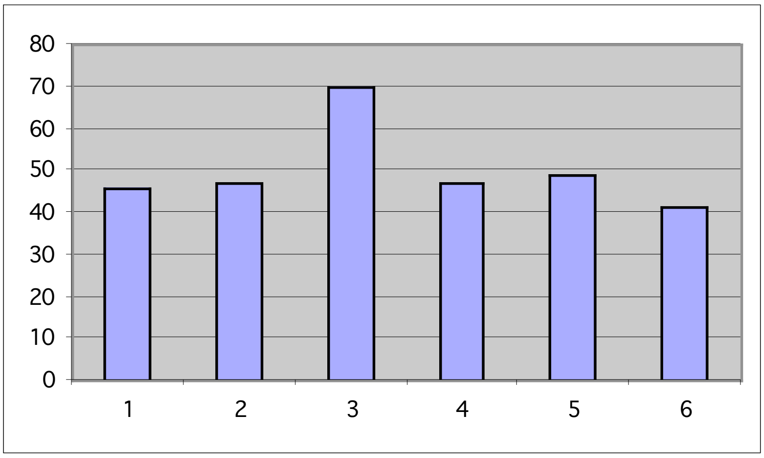
On voudrait savoir si un dé est bien équilibré et si par conséquent la probabilité de chaque face est égale soit 1/6.
Pour cela on jette fois le dé 300 fois et on obtient les résultas suivants :

5	1	6	5	4	2	2	2	3	2
6	4	5	5	3	2	5	6	2	1
3	6	4	2	2	3	1	1	2	6
1	6	4	5	1	4	5	5	4	3
5	2	6	2	3	4	5	2	4	3
3	2	3	5	4	5	3	4	3	5
1	6	3	4	5	6	4	2	4	2
2	4	4	2	4	1	1	3	4	5
3	3	4	3	5	6	3	4	6	3
1	6	3	5	6	1	3	1	4	1
5	3	3	5	4	6	3	4	6	3
3	1	3	5	4	6	6	4	3	3
1	5	1	1	1	1	3	2	5	5
1	4	5	4	4	3	6	3	3	5
2	3	3	4	5	3	1	1	5	4
5	4	3	1	4	6	2	1	3	5
1	2	3	2	6	3	2	4	6	6
4	3	3	6	2	5	5	4	3	1
3	5	2	2	4	3	3	4	6	3
3	3	5	6	5	5	3	2	3	6
1	3	3	6	2	2	5	4	4	1
5	2	3	1	4	1	3	1	6	2
6	1	2	6	4	1	4	1	4	2
3	2	2	3	5	2	1	3	3	6
1	3	5	1	6	3	4	3	3	2
5	4	2	2	3	5	3	3	4	3
1	6	2	5	6	2	2	2	5	6
3	1	5	1	2	6	3	3	3	1
4	5	1	2	1	5	6	1	5	2
6	4	6	1	6	6	5	2	1	5

Ainsi on obtient le tableau ci-dessous, donnant le nombre d'observations pour chaque faces.

1	2	3	4	5	6
46	47	70	47	49	41

L’histogramme des fréquences permet de faire l’hypothèse que la répartition n’est pas



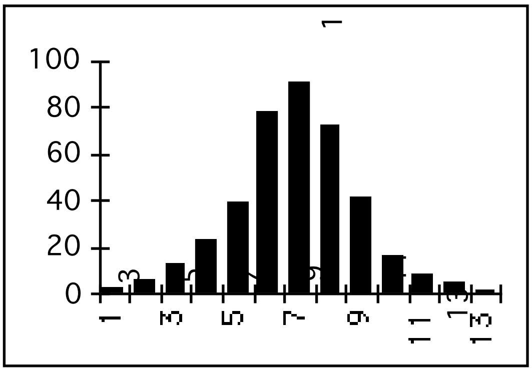
uniforme.

Si on fait l’hypothèse que la répartition est uniforme, la valeur de D est 10,32 et la probabilité pour qu’ une variable aléatoire X suivant une loi de Khi- deux à 5 degrés de liberté dépasse cette valeur est 0,07, l’hypothèse H0 est rejetée, le dé est biaisé.

Exemple 2 cas continu.

Soit un lot de 400 vis dont les unités se répartissent en fonction de leur diamètre suivant les données du tableau ci-dessous :

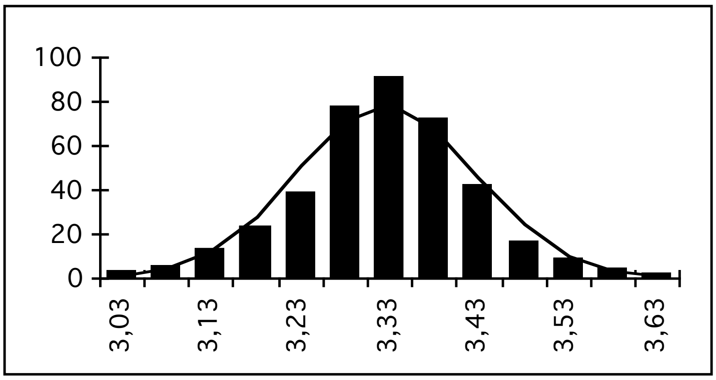
Classes de diamètre (mm)	nombre de vis
3,00 à moins de3,05	3
3,05 à moins de3,10	6
3,10 à moins de3,15	13
3,15 à moins de3,20	23
3,20 à moins de3,25	39
3,25 à moins de3,30	78
3,30 à moins de3,35	91
3,35 à moins de3,40	72
3,40 à moins de3,45	42
3,45 à moins de3,50	17
3,50 à moins de3,55	9
3,55 à moins de3,60	5
3,60 à moins de3,65	2



La forme de l'histogramme suggère l'ajustement à une loi normale. Le calcul de la moyenne et de l'écart -type donne 3,32 et 0,10 mm. On ajustera donc à la distribution la loi normale N(3,32 ;0,10).
Dans l'hypothèse où le diamètre des vis est distribué suivant une loi normale, la probabilité qu'une vis appartienne à une classe est calculée à partir de la table de la loi normale. On obtient un nouveau tableau :

Classes de diamètre (mm)	Probabilité ajustée	Effectif ajusté	nombre de vis
3,00 à moins de3,05	0,0035	1,4	3
3,05 à moins de3,10	0,0104	4,2	6
3,10 à moins de3,15	0,307	12,2	13
3,15 à moins de3,20	0,0707	28,2	23
3,20 à moins de3,25	0,1269	50,8	39
3,25 à moins de3,30	0,1787	71,5	78
3,30 à moins de3,35	0,1972	78,9	91
3,35 à moins de3,40	0,1702	68	72
3,40 à moins de3,45	0,1151	46,1	42
3,45 à moins de3,50	0,0609	24,3	17
3,50 à moins de3,55	0,0252	10,1	9
3,55 à moins de3,60	0,0081	3,3	5
3,60 à moins de3,65	0,0026	1	2

On regroupe les deux classes extrêmes ce qui donne 11 classes, le calcul de d est obtenu à partir du tableau soit d= 12,87. Puisque l'on a estimé la moyenne et l'écart -type, le nombre de degré de liberté est 11-2-1= 8.
La table du Khi Deux donnant la valeur 20 à 1% et 15,5 à 5%, on accepte l'hypothèse de normalité.



EXERCICES DU CHAPITRE V

Exercice 1

Les résultats d'une élection ont donné 40% de voix en faveur d'un candidat. Déterminer rétrospectivement la taille de l'échantillon qui donnerait le résultat avec une précision de ±0,01 avec une probabilité supérieure à 99% :

- En utilisant l'inégalité de Bienaymé – Tchebychev.

- En utilisant l'approximation normale.

Exercice 2

Dans des conditions normales de fonctionnement, la proportion des pièces défectueuses usinées par une machine est égale à 1%.
Un client reçoit une caisse de 100 pièces en provenance directe de la machine. En admettant que celle-ci soit bien réglée, déterminez :

a) La probabilité que le client trouve moins de 2 pièces défectueuses à l'intérieur de la caisse ;

b) La probabilité qu'il y trouve plus de 5 pièces défectueuses.

Exercice 3

Les résultats d'une élection ont donné 45 % des voix en faveur d'un candidat. Déterminez rétrospectivement la probabilité qu'un échantillon aléatoire(de taille N) ait donné la majorité à ce candidat : si N=200, N=1000.

Exercice 4

Une tréfilerie fabrique des câbles métalliques conçus pour résister à de lourdes charges. Leur résistance à la rupture est en moyenne égale à 3 tonnes, avec un écart -type de 200 kg.

Un contrôle de qualité de la fabrication est effectué sur un échantillon de 100 câbles.

En supposant que la production est conforme aux normes de résistance, quelle est la probabilité que la résistance moyenne des câbles de l'échantillon soit comprise entre 2,96 tonnes et 3,02 tonnes ?
Quelle est la probabilité qu'elle soit inférieure à 2,5 tonnes ?

Exercice 5

Deux modèles d'ampoules électriques qualitativement différentes sont produits par une même entreprise.

Les ampoules de type A ont une durée de vie moyenne de 1200h avec un écart type de 200h, les ampoules de type B ont une durée de vie moyenne de 1000h avec un écart type de 100h.
On prélève au hasard un échantillon de 100 ampoules de type A et 150 de type B parmi une production dont nous admettrons la conformité de qualité précédemment énoncées.
Quelle est la probabilité que la durée moyenne de vie observée sur l'échantillon des ampoules de type A soit supérieure de 160h à celle constatée sur l'échantillon d'ampoules de type B ?

Exercice 6

Calculer la probabilité qu'au loto, la boule n°1 sorte entre 30 et 40 fois en 5 ans (259 tirages au total) parmi les 6 numéros ou le complémentaire…

Exercice 7

Une machine débite 1600 cigarettes à la minute. Ces cigarettes ont une masse moyenne de 1,2 gr et un écart type 0,063 gr. A un moment donné on prélève au hasard 30 cigarettes, leur masse moyenne est 1,25 gr.
Est ce que la machine est bien réglée ?

Exercice 8

Une machine automatique est réglée pour remplir des boites de 185 gr en moyenne d'un certain produit. Il arrive parfois que la machine se dérègle. Cependant, même lorsqu'elle fonctionne convenablement, on observe des variations de masse autour de la moyenne.

Sur un échantillon de 30 boites choisies au hasard, on a trouvé une masse moyenne de 183 gr et un écart type de 4 gr. Peut on considérer que la machine est déréglée ? (au seuil de 5%)

Exercice 9

On a tiré un échantillon exhaustif de 10 000 ménages dans une région A comportant 700 000 ménages. Sur cet échantillon on a observé pour un mois déterminé une consommation moyenne d'un certain produit de 950 F avec un écart type de 700 F.

- Calculer l'intervalle de confiance au seuil de 5%, de 1% se rapportant à la moyenne.
-Calculer l'intervalle de confiance au seuil de 5%, de 1% se rapportant à la moyenne si l'on suppose que l'échantillon est indépendant..

-Comparez ces résultats.

Exercice 10

Sur un échantillon de 100 câbles, on a observé une résistance moyenne de rupture de 2,63 tonnes.

a) Sachant que de façon générale la résistance de ce type de câble présente un écart type de 0,025 tonnes, estimer sa résistance moyenne au seuil de 5%

b) Déterminez quelle devrait être la taille minimale de l'échantillon pour connaître la résistance moyenne de ce type de câble à ±1kg au seuil de 5%.

c) On suppose maintenant que l'écart type de la population est inconnu. Un échantillon de 30 câbles indique une résistance moyenne à la rupture égale à 2,66 tonnes avec un écart type de 0,02 tonnes .
Estimez la résistance moyenne au seuil de 5% et déterminez la taille minimale de l'échantillon nécessaire pour connaître cette résistance à ±10kg au seuil de 5%.

Exercice11

Une étude portant sur 30 voitures identiques indique une consommation moyenne de 8,5 litres aux 100 km avec un écart type de 0,8 litres.

a) Donner une estimation de la consommation moyenne de ce type de véhicule au seuil de 5%.

b) Quel devrait être le nombre minimal d'observations auxquelles on devrait procéder pour connaître la consommation moyenne à plus ou moins 2 décilitres près au seuil de 5%, de 1% ?

Exercice 12

On a noté les masses x_i d'un certain nombre d'expédition de pommes(en kg). On a trouvé n_i expéditions de masse x_i

x_i	n_i
95	2
96	1
97	5
98	12
99	21
100	16
101	10
102	7
103	0
104	4

La masse moyenne annoncée des expéditions est de 100kg. Peut on admettre qu'il y ait fraude ? au seuil de 5%)

Exercice 13

Une machine automatique est réglée pour remplir des boites de 185 gr en moyenne d'un certain produit. Il arrive parfois que la machine se dérègle. Cependant, même lorsqu'elle fonctionne convenablement, on observe des variations de masse autour de la moyenne.
Sur un échantillon de 30 boites choisies au hasard, on a trouvé une masse moyenne de 183 gr et un écart type de 4 gr . La machine est elle bien réglée au seuil de 5% ?

Exercice 14
En état normal, une machine débite 1600 cigarettes par minutes. Une cigarette est de masse moyenne de 1, 2 gr et d'écart type 0,063 gr.
A un moment donné on prélève 16 cigarettes à la sortie de la machine leur masse moyenne est de 1,23 gr. Que peut on en conclure ?
Que pourrait on conclure si la même moyenne était constatée sur un échantillon de 90 cigarettes ?

Exercice 15

Un fabricant de matériel électrique a mis au point un nouveau type d'ampoules moins coûteuses. Le fabricant pense que malgré la réduction des coûts de production, les nouvelles ampoules n'ont pas une longévité inférieures à celles des anciennes.(1200 h)

Un échantillon de 100 unités indique une durée de vie moyenne de 1155 h avec un écart type de 50 h .L'opinion du fabricant parait-elle justifiée ? (seuil non précisé)

Exercice 16

On étudie le mouvement propre d'un compteur Geiger -Muller. Le nombre d'impulsions que l'on peut enregistrer, pendant un intervalle de temps donné, est une variable aléatoire dont la valeur moyenne M caractérise le mouvement propre. On a effectué 100 comptages de 10 secondes et on a obtenu :

Nombre d'impulsions	0	1	2	3	4	5	6	7	8
Nombre d'observations	2	8	13	18	20	16	12	8	3

Calculer la valeur moyenne m de cette distribution. Montrer en déterminant l'intervalle de confiance à 95% qu'il est raisonnable d'estimer M par m.

Exercice 17

Une enquête devrait permettre de connaître la surface moyenne des exploitations agricoles d'un pays avec une précision de 5% et un intervalle de confiance à 95%.
L'auteur de l'enquête part précipitamment pour l'étranger ; il n'a pas eu le temps de calculer la taille de l'échantillon permettant d'opérer ce sondage, mais il a laissé' la note suivante : " L'écart -type représente environ la moitié de la moyenne"
Calculer l'effectif de l'échantillon.

Exercice 18

Pour étudier un lot de fabrication de comprimés (lot d'environ 30 000 comprimés) on prélève au hasard 10 comprimés et on les pèse. On observe les 10 valeurs suivantes :

0,81	0,84	0,83	0,80	0,85	0,86	0,85	0,83	0,84	0,80

a) Calculer la moyenne et l'écart type de l'échantillon. Estimer les paramètres de la population totale

b) Quel est l'écart type de l'estimation de la moyenne?

c) On admet que le poids d'un comprimé suit une loi normale, combien faudrait-il prélever de comprimés dans le lot de fabrication étudié pour mesurer le poids moyen d'un comprime avec un écart type de 2 milligrammes?

d) Le poids moyen observé est-il compatible avec la valeur 0,83 g moyenne de la population au seuil de 98%.

Exercice 19

Dans une fabrication de produits chimiques, on s'est aperçu que parmi les f flacons d'un produit P certains contiennent une impureté rendant ce produit inutilisable. On se propose donc de rechercher les flacons défaillants, compte tenu du fait que la probabilité p qu'un flacon contienne l'impureté est indépendante du flacon considéré.

Pour cela, on utilise la méthode suivante : de chacun des flacons, on extrait un échantillon, et les f échantillons ainsi obtenus sont répartis en g groupes de n éléments. Dans chaque groupe, on analyse le mélange des n échantillons. Ainsi, si le mélange ne contient pas d'impuretés, une seule analyse aura établi qu'aucun des n échantillons dans le groupe considéré ne contient d'impuretés. Mais par contre , si le mélange contient l'impureté, il faudra analyser chacun des n flacons (constituant le groupe considéré) séparément afin de déterminer ceux qui sont impurs ; le nombre d'analyses faites pour discriminer ce groupe est alors n+1.

On désigne par X la variable aléatoire égale au nombre d'analyses effectuées pour discriminer les f flacons et on associe à chacun des g groupes (supposés indépendants) l'événement A = " Un échantillon au moins du groupe considéré contient l'impureté".

On désigne par Y la variable aléatoire, égale au nombre de réalisations de l'événement A dans la série des g groupes.

1) Montrer que Y suit une loi Binomiale dont on précisera les paramètres.

2) Exprimer X en fonction de Y.

3) Calculer **E**(X) moyenne de la variable X.

4) Afin de déterminer p, on a choisi au hasard 100 flacons que l'on a analysé. Cette étude préliminaire a donné 1 seul flacon contenant une impureté. Donner au seuil de 5% l'intervalle de confiance se rapportant à p.

5) Déterminer la valeur n_0 de n rendant **E**(X) minimale . Quelle est la valeur ce minimum

6) Application numérique : On suppose que l'on a 1200 flacons à analyser , quelle stratégie doit on mettre en œuvre ?

Exercice 20

Une application directe de la procédure du test au contrôle de la qualité de la fabrication ou des approvisionnements ou de la fiabilité d'un matériel : les cartes de contrôle

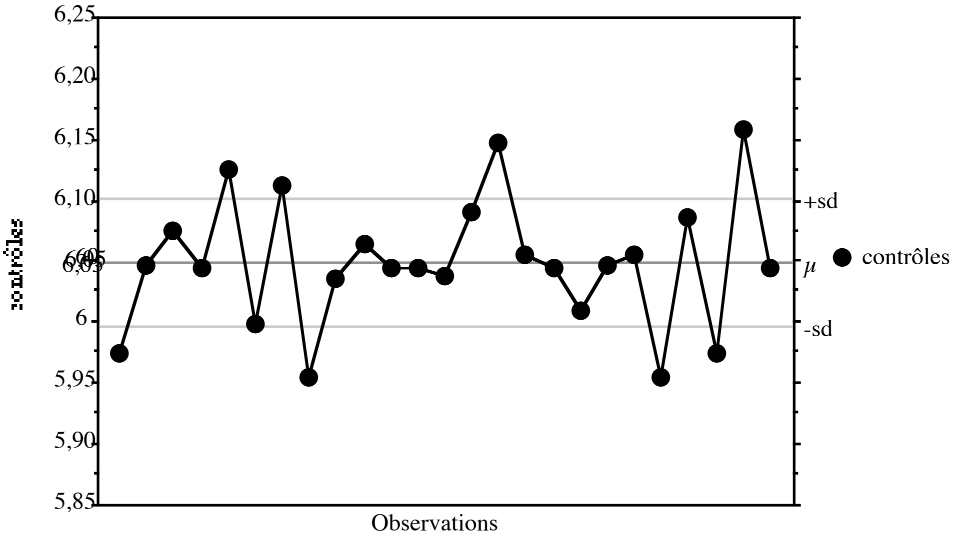
Supposons qu'il s'agisse de contrôler le dosage réalisé automatiquement par une machine, d'une certaine substance (gel filtrant) dans la fabrication d'une crème solaire.

La machine est réglée pour réaliser le dosage conformément à l'indice de protection désiré, soit par exemple, 6,05 ml par tube de 100 ml. Compte tenu de la précision de la machine, on observe dans des conditions normales de fonctionnement, une faible dispersion se traduisant par un écart type de 0,08ml/100ml.

Dans ces conditions, on sait que si l'on tirait au hasard un échantillon de 36 tubes (par exemple),la moyenne des prélèvements serait comprise, au seuil de 5% entre deux valeurs m1 et m2 que l'on calculera. Ces deux limites de la zone d'acceptation (seuil 5%) sont appelées **limites de surveillance** de la moyenne.

Si nous prenons maintenant un seuil de signification de la moyenne beaucoup plus petit, par exemple 0,1%, la moyenne de l'échantillon doit être comprise entre deux valeurs M1 et M2 que l'on calculera. Ces deux limites de la zone d'intervention (seuil 0,1%) sont appelées **limites de contrôle ou d'intervention** de la moyenne.

Concrètement, on procède régulièrement, 5 fois par jour, au prélèvement d'un échantillon de 36 tubes et on calcule la moyenne de chacun de ces échantillons. les résultats sont enregistrés sur un graphique, sur lequel, de part et d'autre de la moyenne théorique m=6ml, on tracera des lignes horizontales correspondant aux limites de surveillance et de contrôle.



Exercice 21

En vue de préciser la configuration des postes de travail dans une chaîne de montage, on a mesuré la distance coude -pouce sur un échantillon de 100 personnes choisies au hasard. On trouve une moyenne égale à 40 cm et un écart -type égal 0,91cm.

a) Peut-on admettre, au seuil de 5% que la distance coude -pouce dans la population est égale à 40,5 cm ?

b) Pour quel seuil de signification le résultat de l'hypothèse s'inverserait-il ?

Exercice 22

En vue d'étudier l'évolution du prix du cuivre aux États-Unis, en Grande-Bretagne et en France, on a calculé pour chacune des 101 années de la période 1892-1992, la valeur du rapport :

$$\frac{P_t}{P_{t-1}} = \frac{\text{prix moyen du cuivre au cours de l' année t}}{\text{prix moyen du cuivre au cours de l' année t - 1}}$$

Aux Etats-Unis, la moyenne de la distribution des 100 rapports successifs ainsi obtenus était égale à 1,004 et l'écart -type de ces rapports était égal à 0,18.

Pour la Grande-Bretagne, on a obtenu, pour la même période, une moyenne de 0,999 et un écart -type de 0,21.

En France, on a observé une moyenne de 0,976 et un écart -type de 0,19.

Au vu de ces résultats, que pensez vous de l'hypothèse selon laquelle le prix de ce métal tendrait à osciller autour d'un niveau approximativement constant (hypothèse de stabilité tendancielle) aux seuils de 5% et de 1% ?

Exercice 23

Pour évaluer la population d'un lac, on pêche une première fois mille poissons, que l'on marque et que l'on rejette dans le lac ; puis on procède à une seconde pêche de mille poissons et on constate que cent parmi eux sont marqués.

a) Soit n le nombre de poissons du lac, calculer la probabilité p(n) pour que, lors de la seconde pêche, il y ait exactement cent poissons marqués.

b) Étudier la suite p(n) et trouver la valeur de n pour laquelle p(n) est maximum.

c) On suppose que lors de la seconde pêche le nombre de poissons marqués suit un loi Binomiale B(1000,p). Donner un intervalle de confiance au seuil de 5 % se rapportant à p. En déduire une estimation du nombre de Poissons du lac et comparer avec les résultats de b).

Exercice 24

L'entreprise Duralumin fabrique un modèle de pièces en inox dont la demande suit approximativement une loi normale dont on ne connaît pas les paramètres.

Pour l'ensemble du marché, on dispose de l'observation de la demande de ces pièces, par semaine, pendant 2 ans, un lot est composé de 100 pièces :

Quantité demandée (en nombre de lots)	**nombre de semaines**
0 à moins de 10	1
10 à moins de 20	2
20 à moins de 30	3
30 à moins de 40	8
40 à moins de 50	25
50 à moins de 60	27
60 à moins de70	20
70 à Moins de 80	12
80 à moins de 90	5
90 et plus	1

1) Donner une estimation ponctuelle des paramètres de la loi normale.

2) Comparer les valeurs théoriques avec les valeurs observées et concluez.

3) L'entreprise établit ses prévisions sur une demande globale annuelle qui ait 60% de chances d'être dépassée. Sa part du marché global est de 1/3.
Déterminer (en nombre de pièces) les ventes que l'entreprise peut espérer réaliser. (On admettra l'indépendance des demandes de chacune des semaines)

<u>Exercice 25</u>

Une machine-outil fabrique des pièces mécaniques dont un certain nombre sont défectueuses et doivent être mises au rebut. On extrait n échantillons de 40 pièces chacun, choisis au hasard dans la fabrication et on fait l'hypothèse que dans la fabrication la proportion p de pièces défectueuses reste constante.

a) Si on note X_i i=1, ..., in les variables aléatoires correspondant au nombre de pièces défectueuses du i- ème échantillon, montrer que $Z = \frac{1}{40n}\sum_{i=1}^{n} X_i$ est un estimateur sans biais de p .

Quelle est la loi de cet estimateur ?

b) Combien faut-il extraire d'échantillons pour obtenir une estimation de p avec une précision de 0,01 au seuil de 5% ?

c) On extrait 100 échantillons de taille 40 et on obtient les résultats suivants:

Nombre de pièces défectueuses	Nombre d'échantillons
0	28
1	40
2	21
3	7
4	3
5	1
6 et plus	0

Donner une estimation de p et un intervalle de confiance de cette estimation au seuil de 5%.

d) Le constructeur de la machine affirme que la proportion p de pièces défectueuses ne dépasse pas 2%.
Que pensez vous de cette affirmation ?