Lois et convergence : un bref résumé synthétique

Loi binomiale
(n tirages oui ou non)

X = 1 si succès probabilité **p**
X = 0 si échec probabilité **q = 1 – p**

Pour X
E(X) = p
V(X) = p – p² = p(1–p) **= pq**

Pour Z = $\sum X_i$ **(nombre de 1 sur n)**

P(Z=k)=C_n^k**p**^{k}**q**$^{n-k}$
E(Z) = np
V(Z) = npq
σ = $\sqrt{npq}$

Pour la fréquence des succès

E($\frac{Z}{n}$**)** =**p** **V(**$\frac{Z}{n}$**)** = $\frac{pq}{n}$

Loi hypergéométrique
(n tirages sans remise)

N boules dont F noires, N-F blanches
X = 1 si noire, n tirages sans remise

Exemple 5 noires , 10 blanches, n=5
P(01010) = $\frac{10}{15}\,\frac{5}{14}\,\frac{9}{13}\,\frac{4}{12}\,\frac{8}{11}$ = P(11000)

P(X=x) = $\frac{C_F^x C_{N-F}^{n-x}}{C_N^n}$

E(X) = p
V(X) = npq $\frac{N-n}{N-1}$

Loi exponentielle
(Durée de vie sans vieillissement)

La probabilité de durer t au-delà de h = probabilité de durer t .

Densité f(t) = $\lambda e^{-\lambda t}$
Répartition P(t <x) = 1 – $e^{-\lambda x}$
Espérance E(t) = $\frac{1}{\lambda}$
Variance V(t) = $\frac{1}{\lambda^2}$
Écart type σ(t)= $\frac{1}{\lambda}$
Médiane $\frac{\ln 2}{\lambda}$

Loi de Poisson
Évènement aléatoire E dans le temps ou l'espace

Si ∆Z = portion de temps ou d'espace
P(E)=p∆Z (P(E) proportionnel à ∆Z)
X =nombre d'événements observés sur Z
Si on pose m = pZ (P(E) sur Z)

P(X=x sur Z) = $\frac{e^{-m}m^x}{x!}$

E(X) = m
V(X) = m

De la loi binomiale à Poisson.
Soit B(n,p) telle que np = 3 ou 4.
Une succession d'épreuves rapides telle que sur ∆t E se produise en moyenne 3 ou 4 fois.
Quand n →∞ et p→ 0 (np → m)

Loi normale
La moyenne (E(x)), la médiane et le mode (P(E) max) coïncident. De plus P(m-2σ < X < m+2σ) > 95%

Densité de probabilité

$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-m)^2}{2\sigma^2}}}$$

Après changement de variable

$\boxed{T = \frac{x-m}{\sigma}}$ **fx dx devient** $\frac{1}{\sqrt{2\pi}} e^{-\frac{T^2}{2}}$ **dT**

Et $\int_{-\infty}^{+\infty} f(T)\,dT$ = 1 d'après Gauss.
Cas général loi N(m,σ) centrée en m
Médiane = m
Mode = m
Écart type = σ
P(T < x) donnée par des tables .
Ne pas oublier que X =Tσ +m

De la loi binomiale à la normale.
Si n est grand et que p proche de ½.
B(n,p) d'écart type σ moyenne np→ N (np,σ)

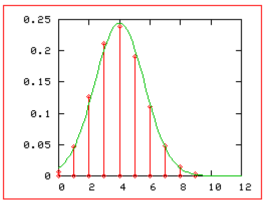


Diagramme en bâtons de B(n,p) et graphe de N(np,σ)

Loi du KHI2 de Pearson

Loi du χ^2 **de Pearson**

On calcule **X=**$\sum_i^n \frac{(Oi-Ei)^2}{Ei}$

Oi =valeur observée Ei = estimée
Le nombre de degrés de liberté est égal au nombre n de données indépendantes diminué du nombre de relations liant ces données. (Par exemple n–1 si on veut évaluer une moyenne ou si la somme des variables = 100).
Dans un tableau destiné à étudier le croisement de L modalités de A et de C modalités de B, le nombre de degrés de libertés est (L–1)(C–1).

La table du khi 2 à k degrés de libertés mous indique quel seuil Q ne doit pas dépasser pour que l’adéquation soit bonne avec telle probabilité. Exemple si Q > 11.07 l’adéquation a 5% de chances d’être bonne. Ce test permet d'estimer l'adéquation de la population à une loi ou à une autre, l'homogénéité ou l'indépendance de 2 populations.

Soit (X_1, X_2,...,X_n) un échantillon de la loi de probabilité de la variable aléatoire X (écart type σ , moyenne m)

Soit la somme: $S_n=X_1+X_2+...+X_n$ On a

E(Sn)=nE(X) = nm	Var(Sn)= nVar(X) = nσ²	σ(Sn)= $\sqrt{n\sigma^2}$ = σ$\sqrt{n}$

Et soit $\bar{X}$ = $\frac{Sn}{n}$ la moyenne

E($\bar{X}$**) = n** $\frac{E(X)}{n}$ **= E(X) = m**	**V(**$\bar{X}$**) =** $\frac{V(X)}{n}$ **=** $\frac{\sigma^2}{n}$	**σ(**$\bar{X}$**) =**$\frac{\sigma}{\sqrt{n}}$

Les formules encadrées en rouge sont importantes car selon qu'on s'intéresse à un nombre d'éléments, une moyenne ou une fréquence on va utiliser l'espérance et l'écart type correspondant pour la convergence vers la loi normale.

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 $\sigma(X) = \sqrt{Npq}$	$E(\bar{X}) = m$ (moyenne ds la population) $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{N}}$	$E(f) = \frac{E(X)}{N} = \frac{Np}{N} = p$ $\sigma(f) = \frac{\sigma(X)}{N} = \frac{\sqrt{Npq}}{N} = \sqrt{\frac{pq}{N}}$

Loi normale N(m,σ)
X suit la loi N(m,σ)
T = $\frac{X-m}{\sigma}$ suit N(0,1)
r% risque de l'estimation

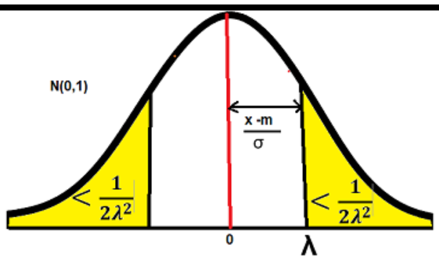
Pour N(0,1) les tables donnent A tel que **P(T∈ [-A, A]) ≥ 1 – r%** en fonction de **1–** $\frac{r}{2}$%.
Pour r =5% cela donne 1– $\frac{r}{2}$% =0,9750 et A =1.96 et donc P(–1,96 < T < 1,96) ≥ 95%
On en déduit qu'**au risque de 5%** **m –1,96σ < X < m +1,96σ**
au risque de 1% **m –2,53σ < X < m +2,53σ**

$$P\left(\frac{|X-m|}{\sigma} \geq \lambda\right) \leq \frac{1}{\lambda^2}$$

Variable centrée
Bienaymé – Tchebychev
P (écart / moyenne) majorée

$$P(|X-m| \geq k) \leq \frac{\sigma^2}{k^2}$$

(k=λσ) Variable quelconque



Comment choisir n de B(n,p) pour que la fréquence du caractère observé soit telle qu'elle ne s'écarte pas plus de 5% de la fréquence effective au risque de 1%. Pour les fréquences m(f) = p σ(f) = $\sqrt{\frac{pq}{n}}$
On a P($\frac{|X-m|}{\sigma} \geq \lambda$ =P(|f – p| ≥σλ) < $\frac{1}{\lambda^2}$ =1%
d'où λ =10 et σλ = 5%= 10 $\sqrt{\frac{pq}{n}}$ d'où on tire n quand on connait p

Grands nombres
Suite Xi de variables de même loi, espérance, variance alors la moyenne de Xi converge en probabilité vers E(x).

m = moyenne dans 'échantillon, S écart type de la moyenne, σ de l'échantillon
Dans P(|m -E(x)|≥ ε) ≤ $\frac{S^2}{\varepsilon^2}$ on remplace S par sa valeur $\frac{\sigma}{\sqrt{n}}$ puis on passe aux complémentaires P(|m -E(x)|≤ ε) ≥ 1– $\frac{\sigma^2}{n\varepsilon^2}$. Cette probabilité tend vers 1 donc quand n grand la moyenne de l'échantillon tend vers E(X) dans la population.
Et la fréquence d'un caractère tend vers sa probabilité dans la population.

Limite centrale.
On prélève n valeurs x1,..,xn de moyenne m tandis que dans la population E(x)=M et σ(X)=σ autrement dit σ(m)=$\frac{\sigma}{\sqrt{n}}$ alors la variable Z=$\frac{m-M}{\sigma(m)}$ tend vers N(0,1)

En particulier B(n,p) converge vers N(np , $\sqrt{npq}$) lorsque n → ∞ (**Moivre – Laplace**).
Par exemple X suit la loi B(500,0,8) moyenne 400 et écart type $\sqrt{80}$
500 est assez grand pour qu'on estime que Y =$\frac{X-400}{\sqrt{80}}$ suit N(0,1)
La moyenne m de X a pour moyenne 0,8 et pour écart type $\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{npq}}{\sqrt{n}}$ = 0,4.
Z = $\frac{m-0{,}8}{0{,}4}$ suit N(0,1) ce qui permet de chercher dans les tables un encadrement de Z au risque de 5% ou 1%. Et d'en déduire un encadrement de m.

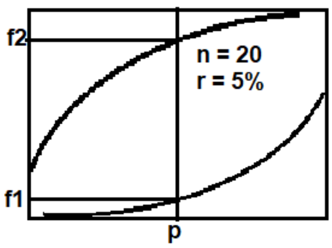
Pour x : M = 800 , σ= 60
Moyenne m de x dans un échantillon de 100 ?

E(m) =800, σ(m) = $\frac{\sigma}{\sqrt{N}} = \frac{60}{10}$ = 6. → $\frac{m-800}{6}$ suit une loi N(0,1)
On peut dire que 800 – 1.96 (6) < m < 800 + 1.96 (6) au risque de 5%.

On connait la fréquence de λ dans la population
P(λ) =p
Fréquence dans un échantillon de taille N ?

On a E(f) = p et σ(f) = $\sqrt{\frac{pq}{N}}$ Si N est grand f suit une loi N(p , $\sqrt{\frac{pq}{N}}$) et donc
Z = $\frac{f-p}{\sqrt{\frac{pq}{N}}}$ suit une loi N(0,1) d'où on peut trouver un encadrement de f au risque de 5%, 1% ou autre.

Autre méthode



Une famille de paires de courbes (ou de tables) paramétrées par la taille N de l'échantillon et le risque de l'encadrement.
On trace la droite x = p (proportion dans la population).
Elle coupe les courbes en 2 points d'ordonnée f1 et f2.
On peut dire que
f1 < f < f2 au risque de r%.

On connait la fréquence et on nous demande la taille de l'échantillon pour précision de 5% au risque de 1%

On a toujours Z = $\frac{f-p}{\sqrt{\frac{pq}{N}}}$ suit une loi N(0,1) .
De l'encadrement –2,58 < Z < +2,58 on déduit p –2,58 $\sqrt{\frac{pq}{N}}$ < f < p +2,58 $\sqrt{\frac{pq}{N}}$
Et 2,58 $\sqrt{\frac{pq}{N}}$ doit être égal à 5% de p (écart maxi toléré) d'où on déduit n .

Loi du khi 2. Adéquation à une loi Homogénéité de 2 popul. Indépendance de 2 popul.	La population est répartie en k classes. Par exemple on lance un dé 90 fois et n(i) est le nombre de fois où on obtient la face i. (6 classes). On calcule $\mathbf{Q} = \sum_{i=1}^{6} \frac{(oi-ei)^2}{ei}$ avec oi valeur observée n(i) et ei effectifs théorique (ici normalement 90 / 6 = 15 pour chaque classe). Le degré de liberté est le nombre de classes diminué du nombre de relation entre les classes. Ici K-1 puisque par exemple n(6)=90-n(1)-n(2)-n3)-n(4)-n(5). Ensuite la table à 5 degrés de liberté nous donne les valeurs V que Q ne doit pas dépasser pour que l'adéquation ait 1 – r% de chances d'être mauvaise , Par exemple si la table donne V =11,07 et P = 0,05 cela signifie que si Q > 11,07 l'adéquation a 95% de chances d'être mauvaise.
Estimateurs T de θ Θ = p, M, σ de la population à partir de n , f, m et V d'un échantillon. Sans biais si lim (T) = θ	Échantillon d'effectif **n** de fréquence **f** , de moyenne **m**, de variance **V**. **f** estimateur sans biais de **p** (proportion du caractère dans la population) **m** estimateur sans biais de **M** (moyenne de la variable aléatoire dans la population) $\mathbf{S^2} = \frac{n}{n-1}$ V est un estimateur sans biais de $\boldsymbol{\sigma^2}$ et donc **S** = $\sqrt{\sum \frac{(Xi-m)^2}{n-1}}$ estimateur de **σ**
Loi de Student Déterminer l'intervalle de confiance de l'espérance E d'une loi normale dont on ne connait pas la variance.	Normalement Z =$\frac{m-E}{\frac{\sigma}{\sqrt{n}}}$ suit une loi normale centrée réduite mais on ne connait pas σ (qui suit une loi du Khi2). La loi de Student donne pour un degré de liberté donné k et pour un risque r le nombre $S^k_{r/2}$ tel que **$P(\lvert Z\rvert > S^k_{r/2}) < r/2$** or cela équivaut (symétrie de la loi et de la courbe densité par rapport à 0) à **$P(m - \frac{\sigma}{\sqrt{n}}S^k_{r/2} \le E \le m + \frac{\sigma}{\sqrt{n}}S^k_{r/2}) > 1-r$** où on remplace σ par son estimateur sans biais S
Comparaison d'une moyenne m à une norme ou un seuil m_0. Si N est assez grand pour une moyenne on utilise N(m_0 , $\frac{\sigma}{\sqrt{N}}$) Faute de connaitre σ on utilise son estimateur sans bais S	On mesure 100 fois une longueur qui devrait être m0. On connait l'écart type σ des mesures m. Au risque de 5% ces mesures ont convenables si $m_0 - 1{,}96\frac{\sigma}{\sqrt{100}} \le m \le m_0 + 1{,}96\frac{\sigma}{\sqrt{100}}$ Si σ est inconnu on le remplace par son estimateur S et on utilise Student à n–1 degrés de liberté pour trouver l'encadrement **$m0 - \frac{\sigma}{\sqrt{n}}S^k_{r/2} \le m \le m0 + \frac{\sigma}{\sqrt{n}}S^k_{r/2}$** Si N grand Student approchée par N (m0 , $\frac{\sigma}{\sqrt{n}}$) On mesure 100 fois x. Peut -on dire que la moyenne m de x ne dépasse pas un seuil m_0 au risque de 1%? On estime σ = S grâce à l'échantillon. Pour N = 100 inutile d'utiliser Student, on utilise N(0,1) pour la variable centrée $\frac{m-m0}{\frac{S}{\sqrt{100}}}$. Hypothèse acceptable si P(m > m0)< 1% soit $\frac{m-m0}{\frac{S}{\sqrt{100}}} \le 2{,}33$ soit $m < m0 + 2{,}33\frac{S}{\sqrt{100}}$
Comparaison d'une fréquence f à une norme un à un seuil p_0 On utilise N(P_0 , $\sqrt{\frac{p0(1-p0)}{N}}$)	N = 100 et le caractère λ est constaté 12 fois (f constatée = 0,12). Peut – on admettre que p = 1/6 au risque de 5% (comparaison à p0 = 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 ; 0,0375) et au risque de 5% on devrait avoir 0,17 – 1,96(0,0375) < f < 0,17 + 1,96(0,0375) soit 0,096 < f < 0,24 et comme f = 0,12 l'hypothèse est admissible.
Comparaison de 2 échantillons.	**Si Xi famille de variables aléatoires indépendantes chacune suivant une loi normale de moyenne mi et d'écart type σi . Alors $\sum Xi$ suit N($\sum mi$, $\sqrt{\sum \sigma i^2}$).**
Tendance et dispersion	On mesure une variable statistique x au sein d'une population. X prend 10 valeurs croissantes de X_0 à X_9 On range les individus selon la valeur de X dans des classes appelées X_0 à X_9 d'effectifs N_0 à N_9 regroupant 100 individus. La moyenne de X est $\sum \frac{XiNi}{100}$ Le mode est la valeur dont la classe a le plus grand effectif par exemple X_3. La médiane C'est la classe Xp dans laquelle l'effectif cumulé N0 +N1 +…+Np atteint la moitié de l'effectif global ici 50 . De la même façon les quartiles, déciles, etc. sont les classes où l'effectif cumulé atteint 1/4, 1/10 de l'effectif global La variance et l'écart type mesurent la dispersion de la série