FACULTÉ DE MÉDECINE
PIERRE & MARIE CURIE

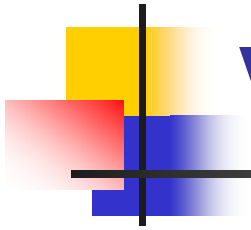
UPMC
SORBONNE UNIVERSITÉS

Probabilités et Biostatistique

2 – Variables aléatoires
Principales lois de probabilité

PAES Faculté de Médecine P. et M. Curie

V. Morice



Variable aléatoire

- Une **variable aléatoire** désigne la grandeur mesurée lors d'une expérience aléatoire
  - Exemples : âge, couleur des yeux
- Résultats possibles de l'expérience $\Rightarrow$ **valeurs possibles** de la variable aléatoire
- Types de variables aléatoires
  - Si résultats numériques (variable **quantitative**)
    - V.a. **continue** : les valeurs couvrent $\mathbb{R}$ ou un intervalle
    - V.a. **discrète** : les valeurs sont discontinues ($\mathbb{N}$)
  - Sinon (variable **qualitative**)
    - V.a. **ordinale** : les valeurs sont ordonnées
    - V.a. **nominale** ou **catégorielle** : valeurs sans ordre

Fonction de répartition

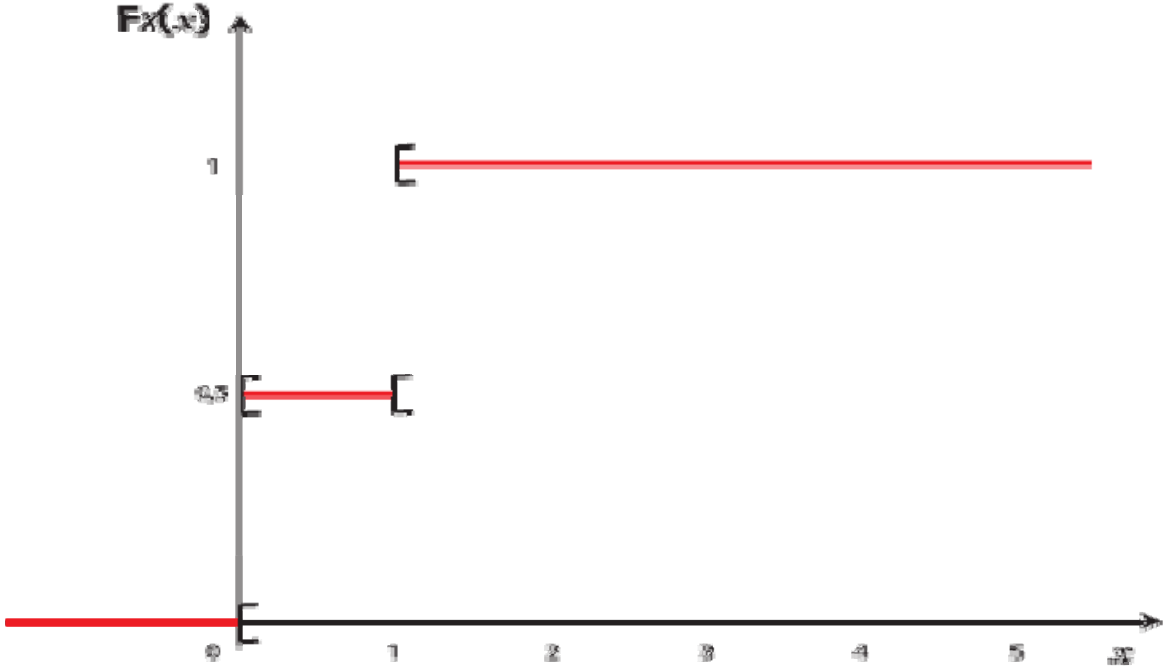
- Soit X une v.a. **quantitative**
- On cherche une fonction définissant la probabilité de tout intervalle [a ; b]
- Soit l'événement $[X \leq x]$ où x est un nombre
- $Pr([X \leq x])$ dépend de la valeur x
- **$F_X(x) = F(x) = Pr([X \leq x])$ = fonction de répartition** de X

Fonction de répartition : premières propriétés

- $F_X(-\infty) = 0$
- $F_X(+\infty) = 1$
- $a < b \Rightarrow$
  $Pr([X \leq b]) = Pr([X \leq a]) + Pr([a < X \leq b])$
  car $[X \leq a]$ et $[a < X \leq b]$ = événements exclusifs
  - $F_X(b) = F_X(a) + Pr([a < X \leq b])$
    - F_X est monotone croissante
    - On trace la courbe en cumulant les probabilités rencontrées lorsque x augmente
  - **$Pr([a < X \leq b]) = F_X(b) - F_X(a)$**

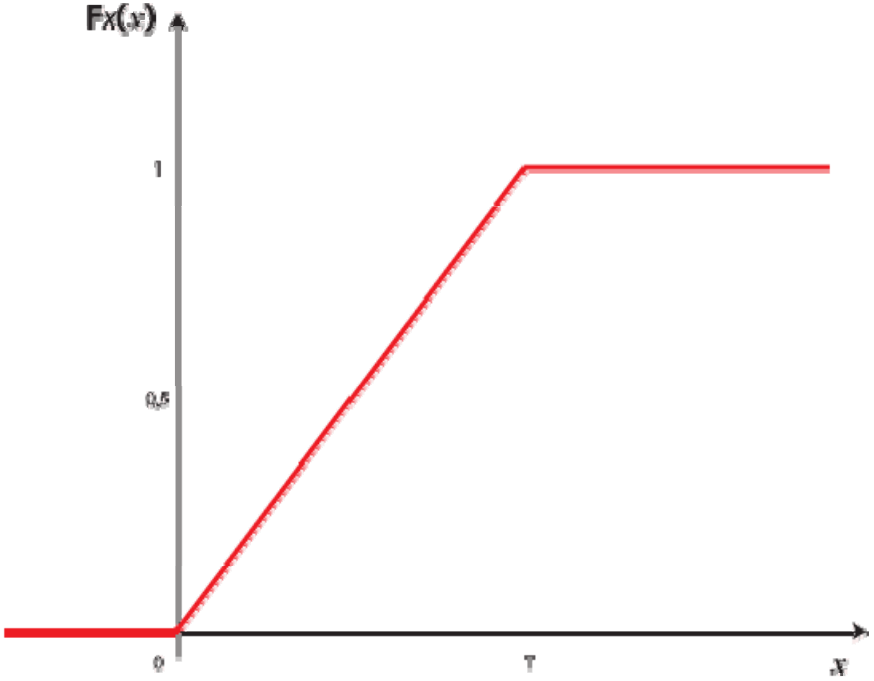
Fonction de répartition : exemple d'une v.a. discrète

- Jet d'une pièce : E = {p, f} ; *Pr* (p) = *Pr* (f) = ½
- V.a. *X* : *X*(f) = 0 ; *X*(p) = 1
- Fonction de répartition



Fonction de répartition : exemple d'une v.a. continue

- Appel téléphonique dans l'intervalle [0,T]
  t =instant d'appel : *Pr* (t1 ≤ t ≤ t2)=(t2-t1)/T (t1 et t2 ∈ [0,T])
- Fonction de répartition

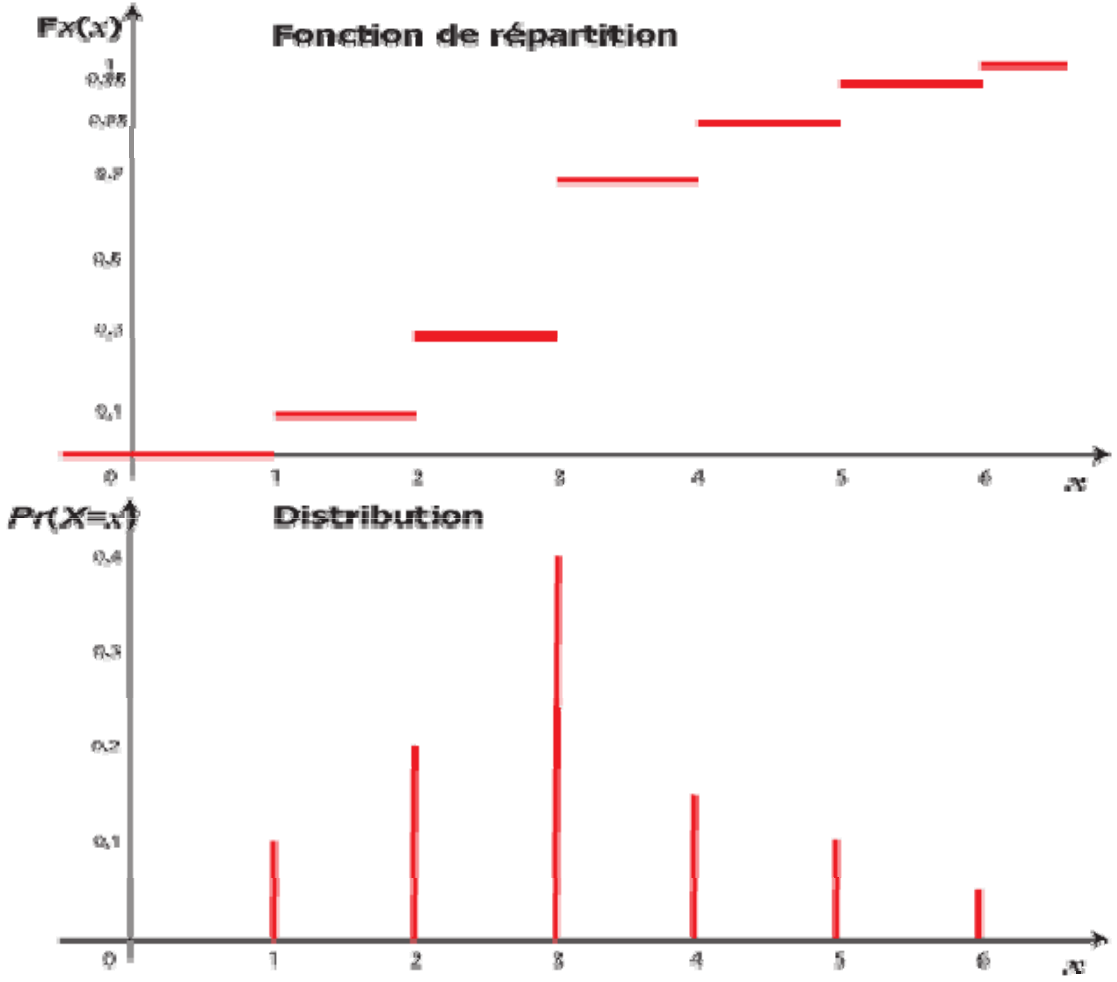


- Si x<0, l'appel n'a pas eu lieu avant x : F(x) = 0
- Si x >T, l'appel a eu lieu avant x : F(x) = 1
- Sinon F(x)=*Pr* (0≤ t≤ x)= x/T

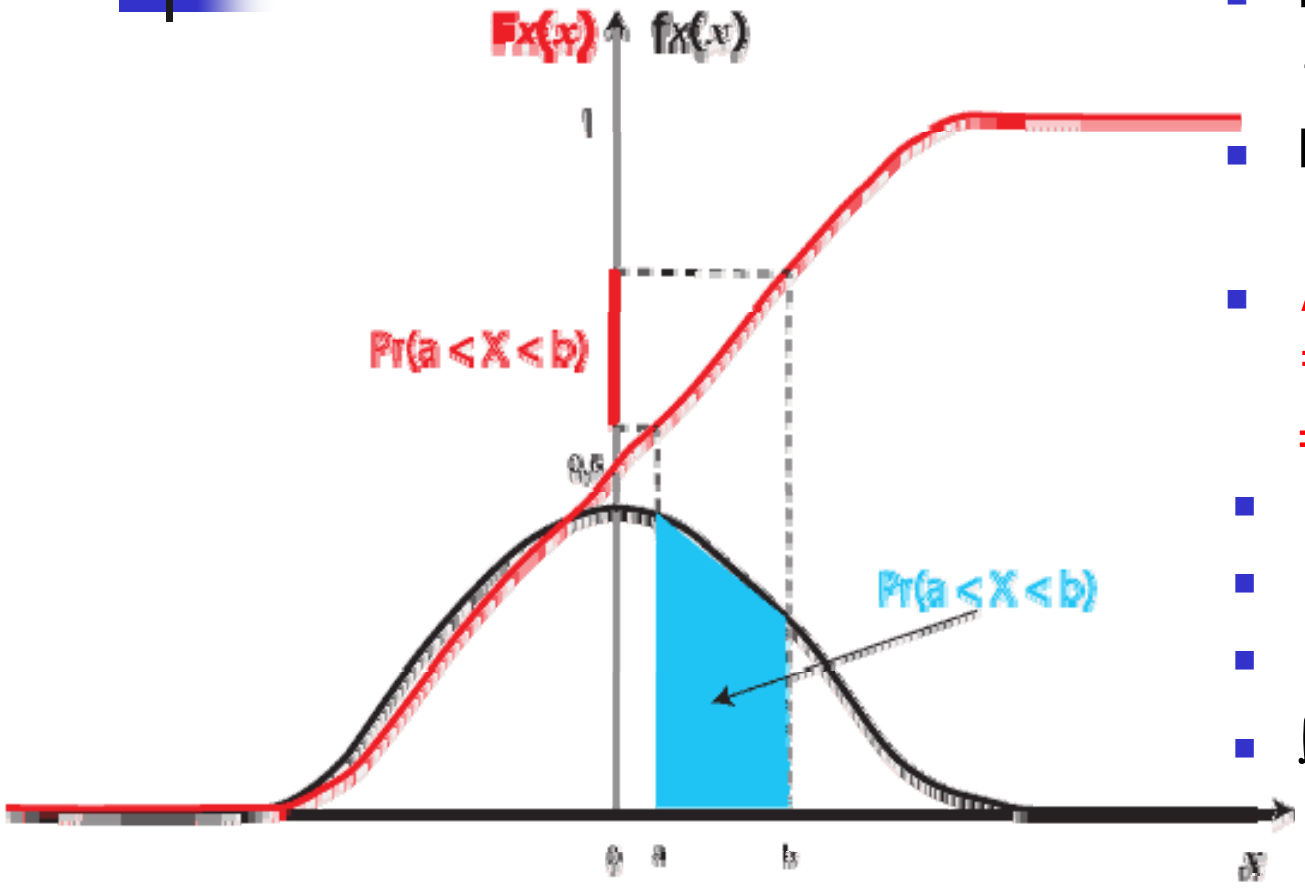
Fonction de répartition : autres propriétés

- On sait $Pr([x^- < X \leq x]) = F_X(x) - F_X(x^-)$
  Si $x^- \to x$, $Pr([x^- < X \leq x]) \to Pr([X = x])$
- Si X est une v.a. continue
  - F_X est continue (si $x^- \to x$, $F_X(x^-) \to F_X(x)$)
  - **Pour tout x, $Pr([X = x]) = 0$**
  - $Pr([a \leq X \leq b]) = Pr([a < X < b])$
- Si X est une v.a. discrète
  - F_X est discontinue
  - En chaque point x de discontinuité, la hauteur du saut ($F_X(x) - F_X(x^-)$ lorsque $x^- \to x$) est la probabilité de x

v.a. discrète : distribution des probabilités



v.a. continue : densité de probabilité



- Densité de probabilité

$$f_X(x) = f(x) = \frac{dF_X(x)}{dx}$$

- Fonction de répartition

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

- $Pr([a \le X \le b]) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$
- $f(x) \ge 0$ (F croissante)
- $f(x)dx = Pr([x \le X \le x+dx])$
- $f(x)dx \approx Pr([X=x])$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Pour définir une v.a. …

	v.a. discrète ou qualitative	**v.a. continue**
Définition de la loi de proba	Tableau des $p_i = Pr(X = x_i)$	Densité de proba f(x) $Pr([a \le X \le b]) = \int_a^b f(x)dx = F(b) - F(a)$
Propriétés	$p_i \ge 0$ $\sum_{i=1}^{n} p_i = 1$ Uniquement si quantitative : $F(x) = \sum_{xi \le x} p_i$	$f(x) \ge 0$ $\int_{-\infty}^{\infty} f(x)dx = 1$ $F(x) = \int_{-\infty}^{x} f(t)dt$ $f(x)dx = Pr(x \le X \le x+dx)$ $f(x)dx \approx Pr(X = x)$

Espérance mathématique [variable quantitative]

- Moyenne au niveau de la **population**
- Notation $E(X) = \mu_X = \mu$
- Calcul : somme de toutes les valeurs pondérées par leur probabilité
  - V.a. discrète : $E(X) = \sum_{i=1}^{n} x_i p_i$
  - V.a. continue : $E(X) = \int_{-\infty}^{\infty} x f(x) dx$

Espérance mathématique : propriétés

Soient des v.a. X et Y et des constantes a, b, c

- $E(c) = c$
- $E(X+c) = E(X)+c$

  Démonstration du cas discret : $Y=X+c$ a pour valeurs $y_i=x_i+c$

  $E(X+c) = E(Y) = \Sigma y_i Pr(Y=y_i) = \Sigma(x_i+c)Pr(Y=y_i)$

  Or $Pr(Y=y_i) = Pr(X+c=x_i+c) = Pr(X = x_i) = p_i$

  Donc $E(X+c) = \Sigma(x_i+c)p_i = \Sigma x_i p_i + c\Sigma p_i = E(X)+c$

  Plus généralement si $Y=g(X)$, on a $\Sigma y_i Pr(Y=y_i) = \Sigma g(x_i)p_i$
- Si $c = -E(X) \Rightarrow E(X-E(X)) = E(X) - E(X) = 0$

  Une v.a. d'espérance nulle est dite centrée
- $E(aX) = aE(X)$
- $E(X+Y) = E(X) + E(Y)$

Variance (et écart-type) [variable quantitative]

- **Variance** = mesure de la variabilité autour de l'espérance
- Notation $\text{var}(X) = \sigma^2_X = \sigma^2$
- Définition **$\text{var}(X) = E[(X-E(X))^2]$**
  On ne peut utiliser $E[X-E(X)]$ qui est nul
- Calcul
  - V.a. discrète $\text{var}(X) = \sum_{i=1}^{n}(x_i-E(X))^2 p_i$
  - V.a. continue $\text{var}(X) = \int_{-\infty}^{\infty}(x-E(X))^2 f(x)dx$
- Autre définition **$\text{var}(X) = E(X^2) - E(X)^2$**
  Car $E[(X-E(X))^2] = E[X^2-2XE(X)+E(X)^2] = E(X^2)-2E(X)E(X)+E(X)^2 = E(X^2) - E(X)^2$
- Calcul
  - V.a. discrète $\text{var}(X) = \sum_{i=1}^{n} x_i^2 p_i - E(X)^2$
  - V.a. continue $\text{var}(X) = \int_{-\infty}^{\infty} x^2 f(x)dx - E(X)^2$
- **Ecart-type** = $\sigma_X = \sigma = \sqrt{\text{var}(X)}$

Variance : propriétés

- $\mathrm{Var}(X) \geq 0$ (somme de carrés)
  Variance nulle pour une constante.
  Variance faible pour une variable peu dispersée
- Si X possède une unité
  - $E(X)$ et σ ont la même unité
  - $\mathrm{Var}(X)$ a cette unité au carré
- Si c est une constante
  - $\mathrm{Var}(c) = 0$
  - $\mathrm{Var}(X+c) = \mathrm{var}(X)$
  - $\mathrm{Var}(c\,X) = c^2\mathrm{var}(X)$
- $\mathrm{Var}(X+Y) = ?$

Loi de 2 variables discrètes ou qualitatives

- X et Y, deux v.a. discrètes ou qualitatives mesurables sur les mêmes individus
- $E_X = \{x_1, x_2, \ldots, x_n\}$; $E_Y = \{y_1, y_2, \ldots, y_m\}$
- Exemple :
  X =sexe (x_1=H ; x_2=F)
  Y =CSP (y_1=agriculteur ; y_2=ouvrier ; … ; y_m=retraité)
- Pour parler simultanément de X et Y, il faut considérer l'espace produit :
  $E_X \times E_Y = \{(x_1,y_1), (x_1,y_2), \ldots, (x_1,y_m), \ldots, (x_n,y_m)\}$
- On doit se donner les probabilités de chaque couple :
  $Pr([X = x_i] \cap [Y = y_j]) = p_{xi,yj}$

Loi de 2 variables discrètes : tableau des probabilités

$X \setminus Y$	y_1	y_2	…	y_m	$\sum_y$
x_1	$p_{x1,y1}$	$p_{x1,y2}$	…	$p_{x1,ym}$	p_{x1}
x_2	$p_{x2,y1}$	$p_{x2,y2}$	…	$p_{x2,ym}$	p_{x2}
…	…	…	…	…	…
x_n	$p_{xn,y1}$	$p_{xn,y2}$	…	$p_{xn,ym}$	p_{xn}
$\sum_x$	p_{y1}	p_{y2}	…	p_{ym}	1

- $p_{xi,yj} = Pr([X = x_i] \cap [Y = y_j])$
- $p_{xi} = \sum p_{xi,yj}$; $p_{yj} = \sum p_{xi,yj}$
- p_x et p_y sont souvent appelées lois marginales
- Ce sont les lois des variables X et Y indépendamment l'une de l'autre

Covariance et corrélation [variables quantitatives]

- $\mathrm{Var}(X+Y) = E[((X+Y)-(\mu_X+\mu_Y))^2] = E[((X-\mu_X)+(Y-\mu_Y))^2]$
  $= E[(X-\mu_X)^2 + (Y-\mu_Y)^2 + 2(X-\mu_X)(Y-\mu_Y)] = \sigma_X^2 + \sigma_Y^2 + 2\mathrm{cov}(X,Y)$
- Première définition : **$\mathrm{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$**
- Seconde définition : **$\mathrm{cov}(X,Y) = E(XY)-\mu_X\mu_Y = E(XY)-E(X)E(Y)$**
  car $E[(X-\mu_X)(Y-\mu_Y)] = E(XY-\mu_X Y-X\mu_Y+\mu_X\mu_Y) = E(XY)-\mu_X\mu_Y-\mu_X\mu_Y+\mu_X\mu_Y$
- Calculs pour deux variables discrètes :
  - $\mathrm{cov}(X,Y) = \sum_{i,j}(x_i-\mu_X)(y_j-\mu_Y)\,p_{xi,yj}$
  - $\mathrm{cov}(X,Y) = \sum_{i,j}x_i y_j\,p_{xi,yj} - \mu_X\mu_Y$
- La covariance est une mesure de l'intensité de la **liaison linéaire** entre deux variables
- Corrélation $\rho_{XY} = \dfrac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}$
- La corrélation est toujours entre -1 et 1

Indépendance de deux variables aléatoires

- X et Y **quantitatives** sont indépendantes si et seulement si les événements $[X \leq x]$ et $[Y \leq y]$ sont indépendants **pour tout x et tout y**
- $\Leftrightarrow$ $\boldsymbol{Pr([X \leq x] \cap [Y \leq y]) = Pr([X \leq x])\, Pr([Y \leq y])}$
- $\Leftrightarrow F_{XY}(x,y) = F_X(x) F_Y(y)$
  où F_X et F_Y sont les fonctions de répartition de X et de Y, et F_{XY} est la fonction de répartition du couple X, Y (définition)
- Si X et Y sont des **v.a. discrètes ou qualitatives**, l'indépendance peut s'écrire (**pour tout x_i et tout y_j**)
  $\boldsymbol{Pr([X = x_i] \cap [Y = y_j]) = Pr([X = x_i])\, Pr([Y = y_j])}$
- $\Leftrightarrow p_{xi,yj} = p_{xi}\, p_{yj}$

Conséquences de l'indépendance de 2 variables quantitatives

Si X et Y sont indépendantes, alors :

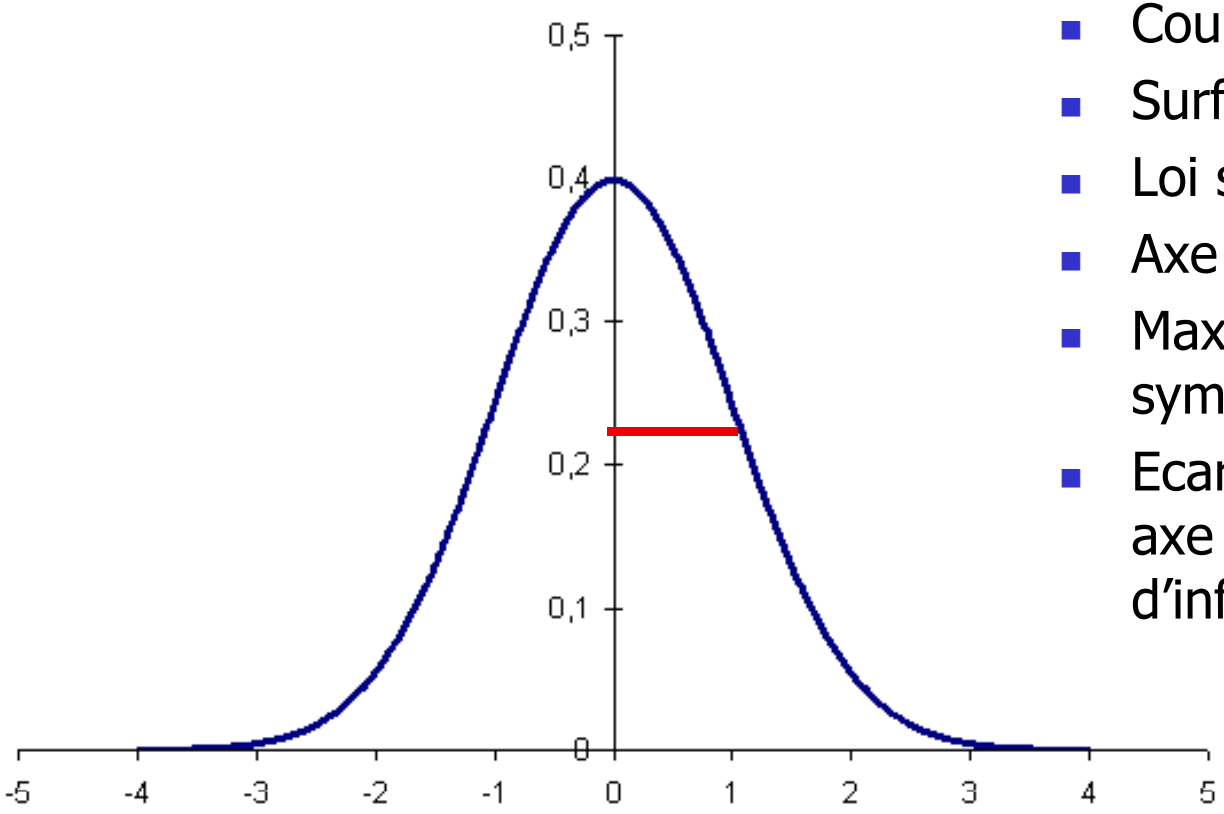
- $\text{cov}(X, Y) = 0$ et $\rho_{XY} = 0$
- $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$
- $E(XY) = E(X)E(Y)$

  car $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$
- La réciproque est fausse

Loi normale $\mathcal{N}(\mu\ ;\ \sigma^2)$

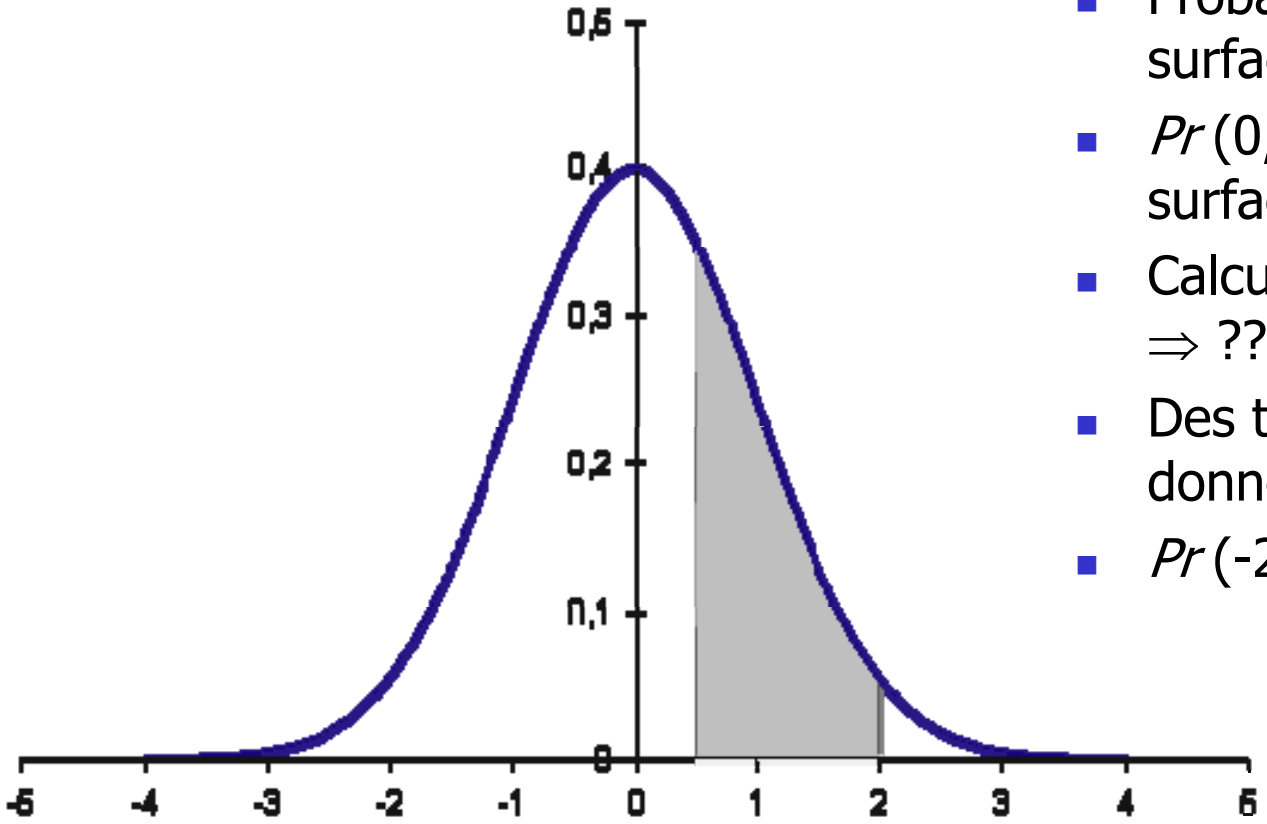
- Loi continue la plus importante
- Densité : $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$
- $E(X) = \mu$
- $\text{var}(X) = \sigma^2$ (donc $\sigma > 0$)
- Si X et Y sont $\mathcal{N}$ et indépendantes, alors $aX+bY$ est $\mathcal{N}$
- Cas particulier $\mathcal{N}(0\ ;\ 1)$
  - Loi centrée ($\mu = 0$) et réduite ($\sigma = 1$)
  - $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Allure de la loi $\mathcal{N}(0 ; 1)$



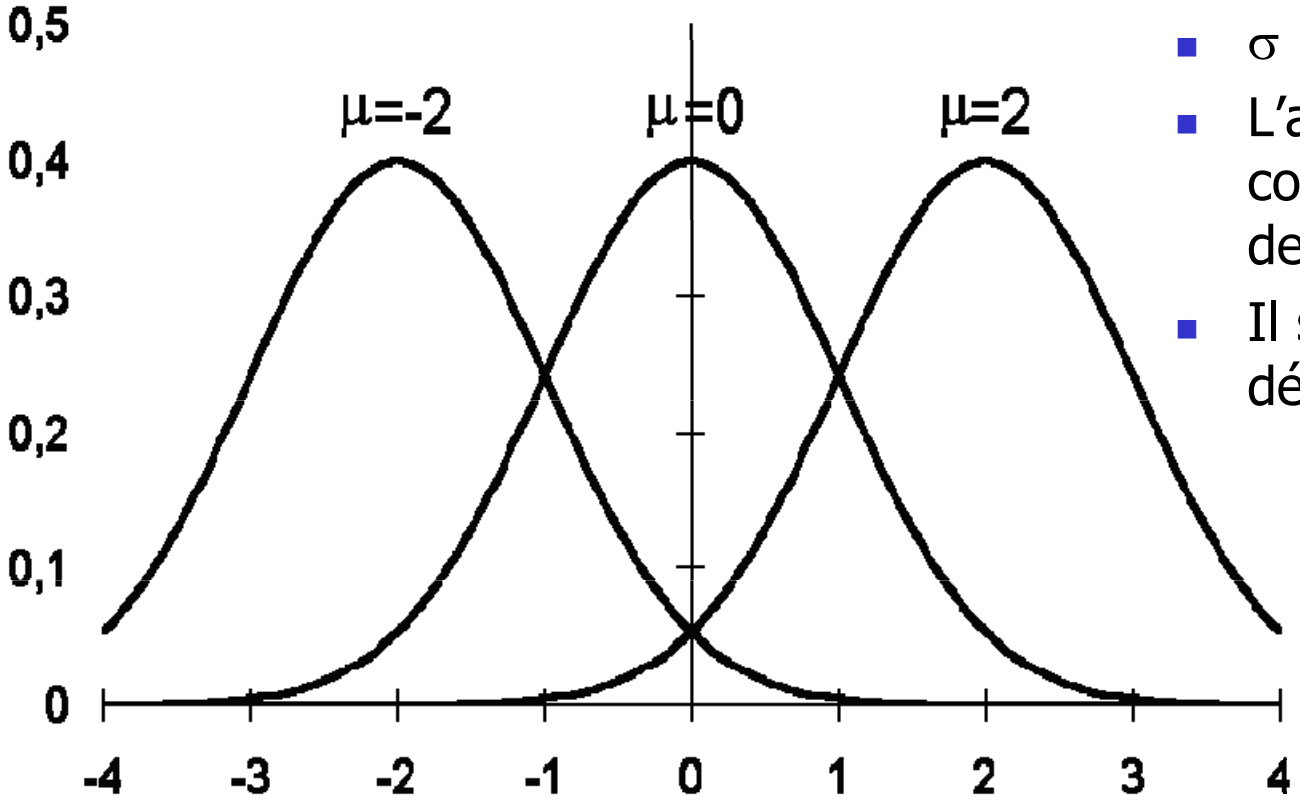
- Courbe de la densité
- Surface sous la courbe = 1
- Loi symétrique
- Axe de symétrie = espérance
- Maximum sur l'axe de symétrie
- Ecart-type = distance entre axe de symétrie et point d'inflexion

Loi $\mathcal{N}(0 ; 1)$ et probabilités



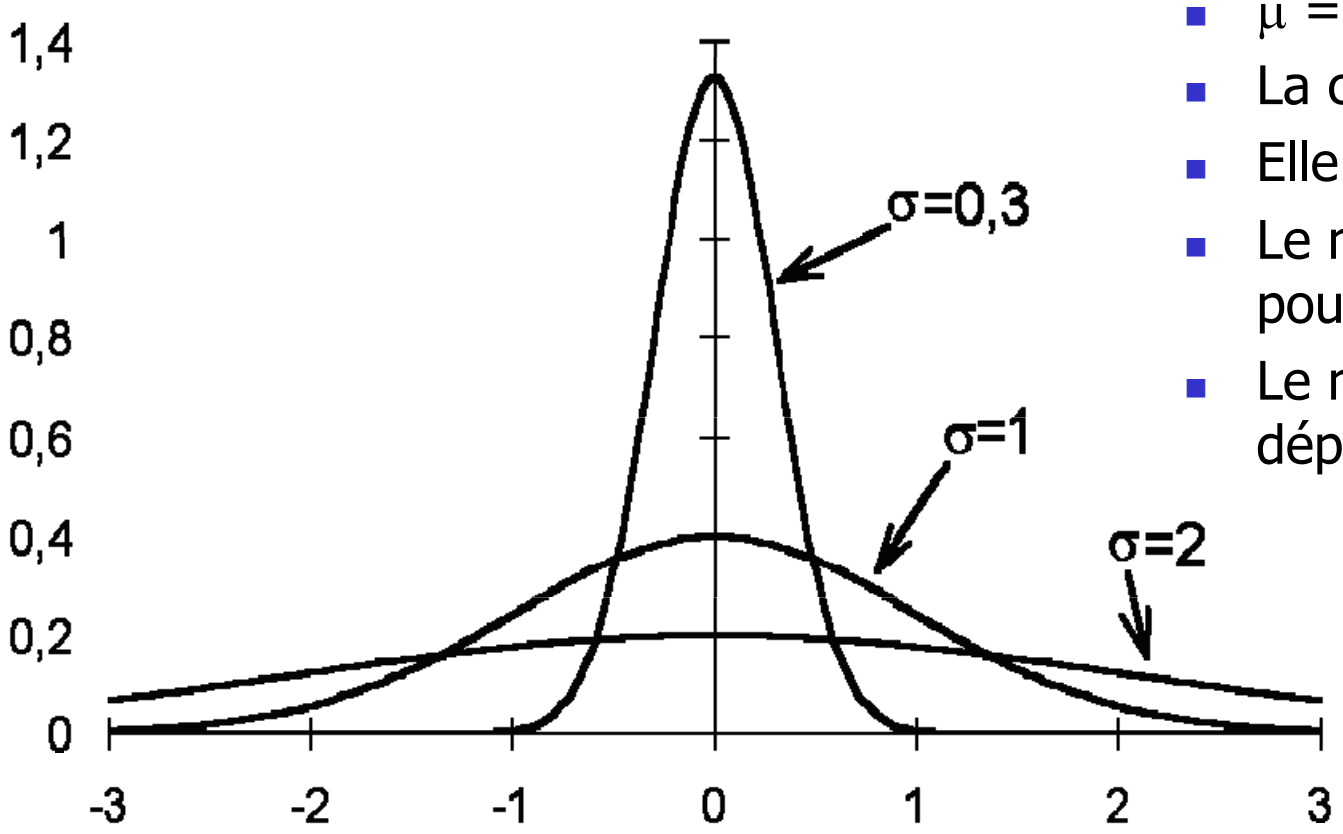
- Probabilité d'un intervalle = surface sous la courbe
- $Pr\,(0{,}5 \leq X \leq 2) = 0{,}312$ = surface grisée
- Calcul = intégration de f(x) $\Rightarrow$???
- Des tables numériques donnent les résultats
- $Pr\,(-2 \leq X \leq 2) \approx 0{,}95$

Loi $\mathcal{N}(\mu ; \sigma^2)$: influence de μ



- σ = 1 pour les 3 courbes
- L'allure de la courbe se conserve si on change de moyenne
- Il s'agit d'un simple décalage

Loi $\mathcal{N}(\mu\ ;\ \sigma^2)$: influence de σ



- $\mu = 0$ pour les 3 courbes
- La courbe s'aplatit si $\sigma\nearrow$
- Elle se resserre si $\sigma\searrow$
- Le maximum s'ajuste pour que la surface = 1
- Le maximum peut dépasser 1

Loi $\mathcal{N}(\mu\ ;\ \sigma^2)$ et probabilités

Soit $X \rightarrow \mathcal{N}(\mu\ ;\ \sigma^2)$. On cherche $Pr\,(a \leq X \leq b)$

- Seule $\mathcal{N}(0\ ;\ 1)$ est tabulée
- Mais $Y = \frac{X-\mu}{\sigma} \rightarrow \mathcal{N}(0\,;1)$
- On va **centrer et réduire** pour obtenir la probabilité

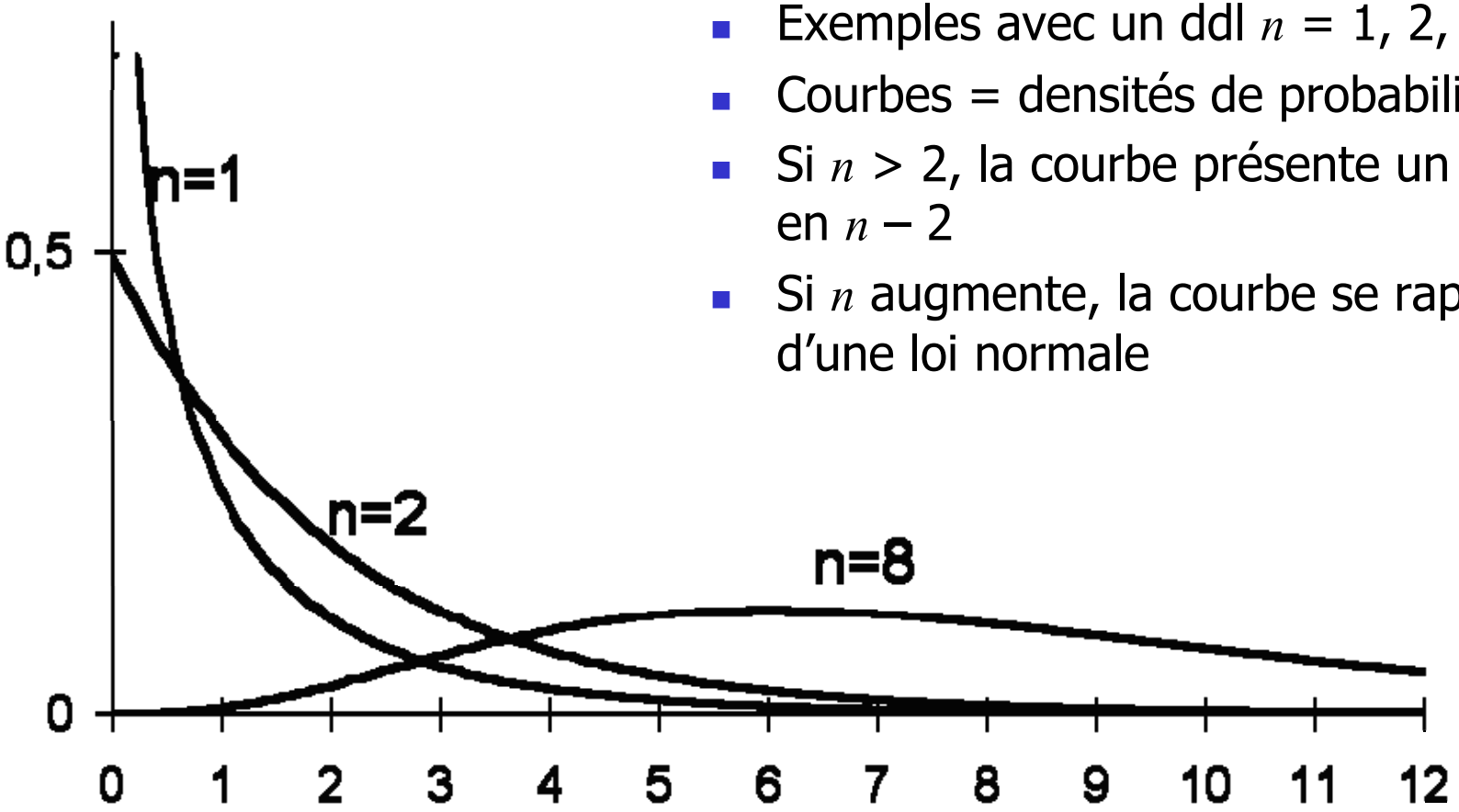
$$Pr(a \leq X \leq b) = Pr(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma})$$

- Posons $c = \frac{a-\mu}{\sigma}$ et $d = \frac{b-\mu}{\sigma}$
- Alors $Pr\,(a \leq X \leq b) = Pr\,(c \leq Y \leq d)$
  La probabilité sur Y se lit dans la table de la loi normale centrée réduite

Loi du « chi-deux » $\chi^2(n)$

- Famille de lois dérivées de $\mathcal{N}(0 ; 1)$
- Si $X_1 \rightarrow \mathcal{N}(0 ; 1)$, alors $X = X_1^2 \rightarrow \chi^2(1)$
- Si $X_1, X_2, \ldots, X_n \rightarrow \mathcal{N}(0 ; 1)$ et sont indépendantes, alors $X = X_1^2 + X_2^2 + \ldots + X_n^2 \rightarrow \chi^2(n)$
- n est le nombre de **degrés de liberté** (ddl)
- $X \geq 0$
- E(X) = n, var(X) = $2n$
- La probabilité d'un intervalle est donnée par une table (qui dépend du ddl)

Allure de la loi du χ^2



- Exemples avec un ddl n = 1, 2, et 8
- Courbes = densités de probabilité
- Si n > 2, la courbe présente un maximum en n – 2
- Si n augmente, la courbe se rapproche d'une loi normale

Loi de Bernoulli

- Base des lois discrètes ou qualitatives
- Expérience à deux résultats possibles *succès* et *échec*
- Variable de Bernoulli : X(échec) = 0, X(succès) = 1
- Pr (succès) = Pr ([X = 1]) = Π
  Pr (échec) = Pr ([X = 0]) = 1 – Π
- E(X) = $\Pi \times 1 + (1 - \Pi) \times 0 = \Pi$
- var(X) = E(X^2) – E(X)^2
  - E(X^2) = $\Pi \times 1^2 + (1 - \Pi) \times 0^2 = \Pi$
  - var(X) = $\Pi - \Pi^2 = \Pi(1 - \Pi)$

Loi binomiale $B(n, \Pi)$

- Construite sur n expériences de Bernoulli **indépendantes** (Π ne change pas entre les épreuves)
- La variable X est le nombre de succès parmi les n expériences (valeur entre 0 et n)
- La probabilité d'avoir exactement k succès est

$$Pr(X=k) = \binom{n}{k} \Pi^k (1-\Pi)^{n-k} = \frac{n!}{k!(n-k)!} \Pi^k (1-\Pi)^{n-k}$$

$\binom{n}{k}$ est le nombre de manières d'obtenir k succès parmi n
$\Pi^k(1-\Pi)^{n-k}$ est la probabilité d'en obtenir une

- $E(X) = n\Pi$; $var(X) = n\Pi(1-\Pi)$

Loi de Poisson

- Loi concernant la réalisation d'événements
  - Faiblement probables (loi des événements rares)
  - Indépendants
  - Exemples : accidents, files d'attente, ruptures de stock
- La variable X est le nombre de réalisations de l'événement
- La loi dépend d'un paramètre λ (λ **> 0**)
- La probabilité d'avoir k réalisations de l'événement rare est

$$Pr(X=k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- Le nombre k de réalisations varie entre 0 et ∞ ($\neq$ loi binomiale)
- $E(X) = \lambda$; $var(X) = \lambda$; $Pr(X=0) = e^{-\lambda}$
- Si $X_1 \rightarrow Poisson(\lambda_1)$, $X_2 \rightarrow Poisson(\lambda_2)$, X_1 et X_2 indépendantes, alors $X=X_1+X_2 \rightarrow Poisson(\lambda_1 + \lambda_2)$

Approximations d'une loi binomiale $B(n, \Pi)$

$$X \rightarrow \boldsymbol{B(n, \Pi)}$$

- Approximation par une loi normale
  - Conditions : $n\Pi \geq 5$ et $n(1\text{-}\Pi) \geq 5$
  - Variable pour l'approximation $Y \rightarrow \mathcal{N}(n\Pi\ ;\ n\Pi(1\text{-}\ \Pi))$
  - On a $Pr([X{=}k]) \approx Pr([k - 0{,}5 \leq Y \leq k + 0{,}5])$
  - Les probabilités $Pr([Y{<}0])$ et $Pr([Y > n])$ sont faibles, mais non nulles
- Approximation par une loi de Poisson
  - Conditions : $\Pi < 0{,}1$ et $n \geq 50$
  - Variable pour l'approximation $Y \rightarrow$ Poisson($\lambda = n\Pi$)
  - On a $Pr([X{=}k]) \approx Pr([Y{=}k])$
  - La probabilité $Pr([Y > n])$ est faible, mais non nulle

Approximation d'une loi de poisson par une loi normale

- $X \rightarrow$ Poisson(λ)
- Conditions : $\lambda > 25$
- Variable pour l'approximation
  $Y \rightarrow \mathcal{N}(\lambda ; \lambda)$
- On a $Pr([X=k]) \approx Pr([k - 0,5 \leq Y \leq k + 0,5])$

Loi de Poisson et risque sanitaire pas encore observé

- Après 10.000 prescriptions d'un nouveau médicament, pas d'effet indésirable
- Que se passera-t-il après 1.000.000 prescriptions ?
- Π = risque individuel d'effet indésirable, inconnu mais faible
- Sur *n* individus, si X est le nombre d'effets indésirables observés, $X \rightarrow B(n, \Pi)$
  - Π faible, *n* grand : $X \rightarrow \text{Poisson}(\lambda = n\Pi)$
  - $\Pr(X=0) = e^{-\lambda} = e^{-n\Pi}$

Loi de Poisson et risque sanitaire pas encore observé (2)

- Que peut-on dire de Π qui soit compatible avec la non observation d'effet indésirable sur *n* individus ?
- Règle : il n'est **pas raisonnable** d'imaginer ne pas observer d'effet indésirable si la probabilité de cette non observation est inférieure à 5%
- Si X=0 sur *n* individus, $Pr(X=0)= e^{-n\Pi} \geq 0,05 \Rightarrow n\Pi \leq 3 \Rightarrow \Pi \leq \mathbf{3/n}$
- La non observation d'effet indésirable sur *n* individus est compatible avec un risque individuel $\Pi \leq 3/n$
- Si n=10000 prescriptions sans effet indésirable, et $\Pi = 3/n = 3 \times 10^{-4}$
  - Avec 1.000.000 de prescriptions on s'attend à 300 effets indésirables
  - Ce qui est énorme