Vocabulaire des statistiques

Une étude porte sur une « **population** » *Par exemple* : Les hommes français de plus de 50 ans. Les sujets atteints du VIH.	**Population** = ensemble des sujets étudiés, décrit avec précision.
Il est souvent impossible d'étudier la grandeur d'intérêt sur toute la « population », on l'observe alors sur une partie seulement de la population, appelée « échantillon » *Par exemple* : les sujets atteints du VIH dépistés en un an dans un centre de santé donné.	**Echantillon**=groupe issu de la population (sous-ensemble de la population). **Echantillon représentatif** : échantillon constitué par tirage au sort dans la population d'étude.

Expériences aléatoires

<table>
<tr><td colspan="2">Résultats non mesurables sur une échelle numérique (modalités) :
l'expérience aléatoire est qualitative</td><td colspan="2">Résultats mesurables sur une échelle numérique : l'expérience aléatoire est quantitative, appelée variable aléatoire</td></tr>
<tr><td colspan="4">Pour savoir si une expérience est qualitative ou quantitative, se poser une question simple : « chaque résultat possible est-il un nombre ? »</td></tr>
<tr><td colspan="2">Deux types d'expériences qualitatives</td><td colspan="2">Deux types d'expériences quantitatives</td></tr>
<tr><td>qualitatives ordonnées

Relation d'ordre évidente entre les résultats possibles</td><td>qualitatives catégoriques

Pas de relation d'ordre évidente entre les résultats possibles</td><td>quantitatives discrètes

L'ensemble des résultats possibles est fini ou dénombrable</td><td>quantitatives continues

L'ensemble des résultats possibles forme un intervalle</td></tr>
<tr><td colspan="2">Pour savoir si une expérience qualitative est ordinale ou catégorique, se poser une question simple : « puis-je ordonner les résultats ? »</td><td colspan="2">Pour savoir si une variable quantitative est discrète ou continue, se poser une question simple : « puis-je numéroter les résultats ? »</td></tr>
<tr><td colspan="4">Remarques :
1. La distinction entre les variables discrètes ou continues dépend de l'échelle de mesure.
En pratique, dès que le nombre de valeurs possibles est élevé, on peut traiter la variable comme continue.
2. Une expérience aléatoire qualitative peut être transformée en variable quantitative discrète en associant des nombres aux résultats.</td></tr>
</table>

Probabilités

Soit E l'ensemble des résultats possibles (ou issues) d'une expérience aléatoire, E est appelé **ensemble fondamental** ou **univers des possibles**.

Evénements

Un événement A est une partie de E.
Les événements peuvent se combiner entre eux pour former de nouveaux événements :

intersection de A et de B :
« A et B »= A ∩B

réunion de A et de B :
« A ou B »=A∪B

contraire de A :
« non A »= $\overline{A}$

Remarquons que $(A \cap B) \cup (A \cap \overline{B}) = A$

Le contraire d'une réunion est l'intersection des contraires :

Non (A ou B) = ni A, ni B
$\overline{A \cup B} = \overline{A} \cap \overline{B}$

Le contraire d'une intersection est la réunion des contraires :

Non (A et B)= Non A ou Non B

$\overline{A \cap B} = \overline{A} \cup \overline{B}$

Probabilité d'un événement

$P(E) = 1 \qquad P(\varnothing) = 0 \qquad 0 \le P(A) \le 1$

$P(\overline{A}) = 1 - P(A)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Evénements incompatibles

A et B sont incompatibles $\Leftrightarrow A \cap B = \varnothing$
$\Leftrightarrow P(A \cap B) = 0$
$\Leftrightarrow P(A \cup B) = P(A) + P(B)$

Formule des probabilités totales

$P(A) = P(A \cap B) + P(A \cap \overline{B})$

En situation d'équiprobabilité

$P(A) = \dfrac{cardA}{cardE}$

Probabilités conditionnelles

$P(A/B) = P_B(A) = \dfrac{P(A \cap B)}{P(B)}$

le contraire de A/B est $\overline{A}/B$

Théorème de la multiplication

$P(A \cap B) = P(A/B) \times P(B) = P(B/A) \times P(A)$

Formule de Bayes

$P(A/B) = \dfrac{P(B/A) \times P(A)}{P(B)}$

Formule des probabilités totales (2ème forme)

$P(A) = P(A/B) \times P(B) + P(A/\overline{B}) \times P(\overline{B})$

Evénements indépendants

A et B sont indépendants $\Leftrightarrow P(A/B) = P(A)$
$\Leftrightarrow P(B/A) = P(B)$
$\Leftrightarrow P(A/\overline{B}) = P(A)$
$\Leftrightarrow P(B/\overline{A}) = P(B)$
$\Leftrightarrow P(A/B) = P(A/\overline{B})$
$\Leftrightarrow P(B/A) = P(B/\overline{A})$

A et B sont indépendants $\Leftrightarrow P(A \cap B) = P(A) \times P(B)$

Attention ! ne pas confondre « indépendants » et « incompatibles »

Signe et diagnostic

Comment faire un diagnostic ?
A partir d'un signe, par exemple une douleur ou le résultat d'un examen.

$S = signe\ présent \quad \overline{S} = signe\ absent$

$T^+ = test\ positif \quad T^- = test\ négatif$

Si le test consiste à mesurer une grandeur continue, une valeur « seuil » est fixée :

Pour une maladie qui augmente la valeur de X	Pour une maladie qui diminue la valeur de X
$Si\ X \geq seuil\ alors\ T^+ \quad Si\ X < seuil\ alors\ T^-$	$Si\ X \leq seuil\ alors\ T^+ \quad Si\ X > seuil\ alors\ T^-$

Les paramètres d'un signe, d'un test diagnostique

$$Sensibilité = Se = P(S/M) = P(T^+/M)$$

$$Spécificité = Sp = P(\overline{S}/\overline{M}) = P(T^-/\overline{M})$$

- Quand on fait varier le seuil, Se et Sp varient en sens inverse.
- Se et Sp ne dépendent pas de la prévalence p de la maladie.

Un sujet peut être :
VP = $T^+ et\ malade$ **FP** = $T^+ et\ sain$ **VN** = $T^- et\ sain$ **FN** = $T^- et\ malade$

Taux de dépistage
« Etre dépisté » signifie « présenter le signe » ou « avoir un test positif »

$$P(T^+) = P(VP) + P(FP)$$
$$P(T^+) = P(T^+ \cap M) + P(T^+ \cap \overline{M})$$
$$P(T^+) = P(T^+/M) \times P(M) + P(T^+/\overline{M}) \times P(\overline{M})$$
$$P(T^+) = p \times Se + (1-p) \times (1-Sp)$$

Probabilité d'erreur

$$P(erreur) = P(FP) + P(FN)$$
$$P(erreur) = P(T^+ \cap \overline{M}) + P(T^- \cap M)$$
$$P(erreur) = P(T^+/\overline{M}) \times P(\overline{M}) + P(T^-/M) \times P(M)$$
$$P(erreur) = (1-p) \times (1-Sp) + p(1-Se)$$

Valeurs prédictives

$$VPP = P(M/T^+) = \frac{P(T^+/M) \times P(M)}{P(T^+)} = \frac{pSe}{pSe + (1-p)(1-Sp)}$$

$$VPN = P(\overline{M}/T^-) = \frac{P(T^-/\overline{M}) \times P(\overline{M})}{P(T^-)} = \frac{(1-p)Sp}{(1-p)Sp + p(1-Se)}$$

- VPP et VPN dépendent de la prévalence p.
- Quand p ↗, la VPP ↗ et la VPN ↘ et inversement.

Variables aléatoires

Une variable aléatoire est une fonction à valeurs dans l'ensemble des réels qui transforme tout événement élémentaire en un nombre (ou observation).
Par extension, un phénomène aléatoire quantitatif est appelé « variable aléatoire »
Soit X une variable aléatoire et E l'ensemble de ses valeurs.

Variable aléatoire discrète

X est **discrète** si E est fini ou dénombrable

$E = \{x_1, x_2, x_3, ..., x_i, ... \}$

Loi de probabilité de X (ou distribution de X) :
ensemble des probabilités des valeurs prises par X

x_1	x_2	...	x_i	...
p_1	p_2		p_i	

$P(X = x_i) =$

$$\sum_i P(X = x_i) = \sum_i p_i = 1$$

Variable aléatoire continue

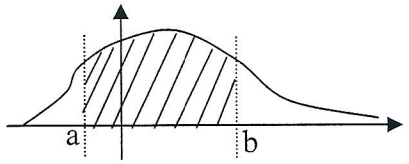
X est **continue** si E n'est pas dénombrable (en général, c'est un intervalle)

La **distribution** de X est déterminée à partir d'une fonction qui est la **densité de probabilité de X.**

Définition : Une fonction f est une densité de probabilité si :

$$\forall x \in \mathbb{R}, f(x) \geq 0 \; et \int_{-\infty}^{+\infty} f(x)dx = 1$$

Autrement dit : si la courbe représentative de f est située au-dessus de l'axe des abscisses et l'aire sous cette courbe est égale à 1. On a alors $P(a < X < b) = \int_a^b f(t)dt$ et $P(X = a) = 0$



Remarque : changer < en ≤ ou > en ≥ ne change pas la probabilité.

Fonction de répartition

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i)$$

F est une fonction en escalier, croissante

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

F est continue, croissante
F est une primitive de la densité.

Espérance (ou moyenne théorique ou moyenne vraie)

$$\mu_X = E(X) = \sum_i x_i P(X = x_i) = \sum_i p_i x_i$$

$$\mu_X = E(X) = \int_{-\infty}^{+\infty} x f(x)dx$$

Variance (vraie ou théorique) et écart-type

$$Var(X) = E(X - E(X))^2 = E(X^2) - (E(X))^2$$

$$\sigma_X = \sqrt{Var(X)}$$

$$Var(X) = \sum_i p_i (x_i - E(X))^2$$

$$Var(X) = \sum p_i x_i^2 - (E(X))^2$$

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 f(x)dx$$

$$Var(X) = \int_{-\infty}^{+\infty} x^2 f(x)dx - (E(X))^2$$

Règles de calcul avec les espérances et les variances

X et Y sont des variables aléatoires discrètes ou continues, a est une constante réelle

L'espérance est linéaire

$E(X+Y)=E(X)+E(Y)$
$E(aX)=aE(X)$
$E(a)=a$
$E(X+a)=E(X)+a$

$Var(aX)= a^2 Var(X)$

$Var(a)=0$
$Var(X+a)=Var(X)$

Si X et Y sont INDEPENDANTES ce qui signifie :

$$P((X=x_i)\cap(Y=y_j))=P(X=x_i)\times P(Y=y_j) \text{ pour tous les couples } (i,j)$$

Alors :

Var(X+Y)=Var(X)+Var(Y)

La réciproque est fausse.

Lois discrètes usuelles à connaître.

Loi de Bernoulli de paramètre π	Expérience aléatoire à deux résultats possibles (épreuve de Bernoulli) X=1 si succès X=0 si échec P(X=1)= π P(X=0)=1-π	$E(X)=\pi$ $Var(X)=\pi(1-\pi)$
Loi binomiale de paramètres n et π	La répétition n fois de manière indépendante de la même épreuve de Bernoulli définit n variables de Bernoulli $X_1, X_2, \ldots, X_n$ de même paramètre π La variable aléatoire S_n égale au nombre de succès sur les n résultats prend ses valeurs dans $\{0,1,2.....,n\}$ et est égale à : $S_n= X_1+X_2+\ldots+X_n$ $P(S_n=k)=C_n^k\pi^k(1-\pi)^{n-k}$ $C_n^k=\binom{n}{k}=\frac{n!}{k!(n-k)!}$	$E(S_n)=n\pi$ $Var(S_n)=n\pi(1-\pi)$
Loi de Poisson de paramètre λ (λ est un réel strictement positif)	X= nombre d'événements observés pendant une période de temps donnée dans le cas où ces événements sont indépendants et faiblement probables (nombre de colonies bactériennes dans une boîte de Petri, nombre d'accidents …) X prend ses valeurs dans ℕ $P(X=k)=e^{-\lambda}\frac{\lambda^k}{k!}$ *Remarque : Si X_1 et X_2 sont **indépendantes** et distribuées selon des lois de **Poisson** de paramètres λ_1 et λ_2 alors X_1+X_2 suit une loi de **Poisson** de paramètre $\lambda_1+\lambda_2$*	$E(X)=Var(X)=\lambda$

Approximation de la loi binomiale par la loi de Poisson
Si X suit une loi **binomiale** de paramètres n et π avec ***n* grand (*n*>50) et π petit (π<0,1)** alors X suit **approximativement** une loi de **Poisson** de paramètre $n\pi$

Lois continues usuelles à connaître

Loi uniforme sur [a ;b] *modélise les tirages de nombres au hasard dans un intervalle.*	E=[a ;b] Densité de probabilité $\begin{cases} f(x)=\dfrac{1}{b-a} \; si \; x\in[a;b] \\ f(x)=0 \; si \; x\notin[a;b] \end{cases}$	$E(X)=\dfrac{a+b}{2}$ $Var(X)=\dfrac{(b-a)^2}{12}$
Loi exponentielle de paramètre λ>0 *modélise les durées de vie lorsque le vieillissement n'intervient pas.*	E=$\mathbb{R}^+$ Densité de probabilité $\begin{cases} f(x)=\lambda e^{-\lambda x} \; si \; x\geq 0 \\ f(x)=0 \; si \; x<0 \end{cases}$ Fonction de répartition : Pour $x\geq 0$, $F(x)=P(X\leq x)=1-e^{-\lambda x}$ **Propriété importante :** Si X suit une loi exponentielle Alors $P(X>a+b/X>a)=P(X>b)$ La loi exponentielle est dite « *sans mémoire* ».	$E(X)=\dfrac{1}{\lambda}$ $Var(X)=\dfrac{1}{\lambda^2}$ $med(X)=\dfrac{\ln 2}{\lambda}$
Loi normale (ou de Gauss) de paramètres μ et σ>0 **notée** $N(\mu\,;\sigma^2)$ *modélise de nombreuses mesures biologiques.*	E=$\mathbb{R}$ Densité de probabilité $f(x)=\dfrac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ la courbe de la densité est une courbe en cloche symétrique par rapport à la droite d'équation $x=\mu$ *Dans la pratique :* Si X suit $N(\mu\,;\sigma^2)$ Alors $Z=\dfrac{X-\mu}{\sigma}$ suit $N(0\,;1)$ *Autrement dit :* $Z=\dfrac{X-\mu}{\sigma}$ est normale **centrée réduite** *puis on utilise la **table** de la loi normale centrée réduite pour le calcul des probabilités*	$E(X)=\mu$ $Var(X)=\sigma^2$

Intervalles de pari d'une variable aléatoire Normale X de moyenne E(X)=μ et d'écart-type σ

L'intervalle de **pari** (appelé aussi intervalle de **fluctuation**) de X (Normale) de **niveau** $1-\alpha$ (ou de **risque** α) est :
$[\mu-\varepsilon_\alpha\sigma\ ;\ \mu+\varepsilon_\alpha\sigma]$ C'est l'intervalle $[a,b]$ **centré sur** μ tel que $P(a \le X \le b) = 1-\alpha$

Cas particulier classique : L'IP de X de niveau 95% (de risque 5 %) est $[\mu-2\sigma\ ;\ \mu+2\sigma]$ (avec l'arrondi habituel de 1,96 à 2)
95% des réalisations de X sont dans cet IP
Il y a 95 chances sur 100 que X prenne sa valeur dans cet IP
Remarque : Il existe d'autres intervalles que l'IP de niveau $1-\alpha$ (**non centrés sur** μ) ayant pour probabilité $1-\alpha$

Variable aléatoire Somme, variable aléatoire Moyenne, variable aléatoire Proportion

Soit *n* réalisations d'une variable aléatoire X, par exemple la mesure d'une grandeur biologique chez *n* sujets (variable aléatoire continue) ou l'étude d'une propriété en codant 1 la présence de cette propriété et 0 son absence (variable aléatoire de Bernoulli).

Ces *n* répétitions indépendantes de X sur *n* sujets sont *n* variables aléatoires $X_1, X_2, \ldots, X_n$ de même loi que X (**iid** : indépendantes identiquement distribuées), donc de moyenne (espérance) μ et d'écart-type σ.

A partir de *n* réalisations de X (n valeurs prises par X) dans un échantillon de *n* sujets, on peut en calculer la somme qui est la valeur prise par **la variable aléatoire Somme** $S_n = \sum_{i=1}^{n} X_i$

ou la moyenne arithmétique (moyenne observée) qui est la valeur prise par **la variable aléatoire Moyenne**

$M_n = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{S_n}{n}$

$E(S_n) = nE(X) = n\mu$ $\quad Var(S_n) = nVar(X)$ $\quad$ écart type.

$E(M_n) = E(X) = \mu$ $\quad Var(M_n) = \frac{Var(X)}{n}$ $\quad \sigma(M_n) = \frac{\sigma(X)}{\sqrt{n}}$

Remarque importante :
Dans le cas où X est une **variable aléatoire de Bernoulli**,

$M_n = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{S_n}{n} = \frac{\text{v.a. Nombre de succès dans un échantillon de taille } n}{n}$

= *v.a. Proportion de succès dans un échantillon de taille n*

M_n *est alors notée alors* P_n

La variable aléatoire Proportion P_n est la variable aléatoire Moyenne M_n de *n* répétitions d'une variable de Bernoulli.

Ne pas confondre P_n et S_n : $P_n = \frac{S_n}{n}$ **S_n suit une loi binomiale, mais pas P_n !**

$E(P_n) = E(X) = \pi$ $\quad Var(P_n) = \frac{Var(X)}{n} = \frac{\pi(1-\pi)}{n}$ $\quad \sigma(P_n) = \sqrt{\frac{\pi(1-\pi)}{n}}$

Loi de S_n , Loi de M_n , Loi de P_n

Théorème central limite : Quelle que soit la loi de X, variable aléatoire de moyenne μ et d'écart-type σ, les lois de S_n et de M_n (donc aussi de P_n) convergent, pour n suffisamment grand, vers une loi normale.

Plus précisément :

X ≠ Bernoulli	Si $X \sim N(\mu\ ;\ \sigma^2)$ ***pas besoin du TCL*** ⟶	$S_n \sim N(n\mu\ ;\ n\sigma^2)$ (quel que soit n) **Loi exacte**	$M_n \sim N(\mu\ ;\ \frac{\sigma^2}{n})$ (quel que soit n) **Loi exacte**
	Si $X \sim Poisson(\lambda)$ ***pas besoin du TCL*** ⟶	$S_n \sim Poisson(n\lambda)$ (quel que soit n) **Loi exacte**	M_n **Loi exacte inconnue**
	Si de plus $\boxed{n \geq 30}$ **TCL** ⟶	$S_n \sim N(n\lambda\ ;\ n\lambda)$ ***Loi approchée***	$M_n \sim N(\lambda\ ;\ \frac{\lambda}{n})$ ***Loi approchée***
	Si X ne suit pas une loi Normale (autre loi ou loi inconnue, moy μ et var σ^2) et $\boxed{n \geq 30}$ **TCL** ⟶	$S_n \sim N(n\mu\ ;\ n\sigma^2)$ ***Loi approchée***	$M_n \sim N(\mu\ ;\ \frac{\sigma^2}{n})$ ***Loi approchée***
X de Bernoulli de paramètre π	***pas besoin du TCL*** ⟶	S_n ~ loi binomiale de paramètres (n, π) **Loi exacte**	P_n **Loi exacte inconnue**
	Si $\boxed{n\pi \geq 5 \ \ et \ \ n(1-\pi) \geq 5}$ **TCL** ⟶	$S_n \sim N(n\pi\ ;\ n\pi(1-\pi))$ ***Loi approchée***	$P_n \sim N(\pi; \frac{\pi(1-\pi)}{n})$ ***Loi approchée***

Remarque : On dispose finalement de deux approximations possibles pour une loi **binomiale** :

Approximations de la loi binomiale

1) par la loi de Poisson (voir séance 2)

Si X suit une loi **binomiale** de paramètres n et π

avec ***n* grand (*n*>50) et π petit (π<0,1)**

alors X suit **approximativement** une loi de **Poisson** de paramètre $n\pi$

2) par la loi normale (voir ci-dessus)

Si X suit une loi **binomiale** de paramètres n et π

avec $\boxed{n\pi \geq 5 \ \ et \ \ n(1-\pi) \geq 5}$

alors X suit **approximativement** une loi normale $N(n\pi; n\pi(1-\pi))$

Approximation de la loi de Poisson par la loi Normale:

Si X suit une loi de Poisson avec paramètre grand $\boxed{\lambda > 30}$

alors X suit **approximativement** une loi Normale d'espérance et variance égales à λ.

Remarque : Quand S_n , M_n ou P_n suit une loi normale (exacte ou approchée par le TCL, voir conditions dans tableau), on peut écrire des intervalles de pari de ces variables aléatoires, avec leur propre moyenne et leur propre écart-type bien sûr !

$$IP_{niv\,1-\alpha}(S_n) = [\mu(S_n) - \varepsilon_\alpha \sigma(S_n); \mu(S_n) + \varepsilon_\alpha \sigma(S_n)] = [n\mu_X - \varepsilon_\alpha \sqrt{n}\sigma_X ; n\mu_X + \varepsilon_\alpha \sqrt{n}\sigma_X]$$

$$IP_{niv\,1-\alpha}(M_n) = [\mu(M_n) - \varepsilon_\alpha \sigma(M_n); \mu(M_n) + \varepsilon_\alpha \sigma(M_n)] = [\mu_X - \varepsilon_\alpha \frac{\sigma_X}{\sqrt{n}} ; \mu_X + \varepsilon_\alpha \frac{\sigma_X}{\sqrt{n}}]$$

$$IP_{niv\,1-\alpha}(P_n) = [\mu(P_n) - \varepsilon_\alpha \sigma(P_n); \mu(P_n) + \varepsilon_\alpha \sigma(P_n)] = [\pi - \varepsilon_\alpha \sqrt{\frac{\pi(1-\pi)}{n}} ; \pi + \varepsilon_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}]$$

Estimation

Problème de l'estimation : on cherche à connaître un paramètre θ d'une variable aléatoire (en général la moyenne vraie μ ou la variance vraie σ^2) grâce à des observations réalisées sur un échantillon de taille n extrait de la population.

Estimation ponctuelle d'un paramètre θ

Pour approcher un paramètre **inconnu** θ, on utilise un **estimateur**, c'est-à-dire une variable aléatoire T_n qui prend, dans des échantillons de taille n, des valeurs voisines du paramètre θ que l'on cherche à estimer.
Une valeur prise par T_n est **une estimation ponctuelle** de θ

Biais de l'estimateur T_n : $B(T_n) = E(T_n) - \theta$

L'estimateur T_n est **non biaisé** si son biais est nul, autrement dit si $E(T_n) = \theta$

Dans le cas où T_n est biaisé, si $\lim_{n \to +\infty} B(T_n) = 0$ on dit que l'estimateur T_n est **asymptotiquement non biaisé**

Soit X une variable aléatoire de moyenne vraie μ et de variance vraie σ^2 inconnues	
Estimateur usuel Tous les estima. usuel sont non biaisés	Sur un échantillon de n sujets on peut calculer la valeur prise par l'estimateur
De μ : $M_n = \frac{\sum_{i=1}^{n} X_i}{n}$ non biaisé	$m = \frac{x_1 + x_2 + \ldots + x_n}{n}$ moyenne **observée**, estimation ponctuelle de la moyenne vraie μ.
De π, proportion vraie (paramètre d'une loi de Bernoulli, autrement dit moyenne vraie d'une loi de Bernoulli) : $P_n = \frac{\sum_{i=1}^{n} X_i}{n}$ non biaisé	$p = \frac{nombre\ de\ succès\ dans\ l'échantillon}{n}$ proportion **observée** de succès, estimation ponctuelle de la proportion vraie π .
De σ^2 : $V_n \ ou \ S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - M_n)^2$ $= \frac{1}{n-1}[\sum_{i=1}^{n} X_i^2 - nM_n^2]$ non biaisé	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - n \times m^2]$ variance **observée**, estimation ponctuelle de la variance vraie σ^2.

Remarque : Des échantillons différents fournissent des estimations ponctuelles différentes.

Estimation par intervalle d'une moyenne vraie ou d'une proportion vraie

Il s'agit de donner une fourchette de valeurs pour μ ou π (paramètre inconnu) : un intervalle dans lequel se trouve le paramètre inconnu avec une probabilité fixée $1-\alpha$, appelé **intervalle de confiance de niveau $1-\alpha$ ou de risque α**.

IC d'une moyenne vraie μ (d'une variable aléatoire X ≠ Bernoulli) **établi à partir d'un échantillon de *n* sujets**

On calcule m la moyenne observée et s^2 la variance observée dans l'échantillon

Condition de validité :

Soit X est Normale (alors pas de condition sur *n*)

Soit n ≥ *30*

$$IC_{niv\,1-\alpha}(\mu) = [m - \varepsilon_\alpha \frac{s}{\sqrt{n}} ; m + \varepsilon_\alpha \frac{s}{\sqrt{n}}]$$

précision de l'IC : $\varepsilon_\alpha \frac{s}{\sqrt{n}}$ largeur de l'IC : $2 \times \varepsilon_\alpha \frac{s}{\sqrt{n}}$

IC d'une proportion vraie π **établi à partir d'un échantillon de *n* sujets**

On calcule p la proportion observée

$$IC_{niv\,1-\alpha}(\pi) = [p - \varepsilon_\alpha \sqrt{\frac{p(1-p)}{n}} ; p + \varepsilon_\alpha \sqrt{\frac{p(1-p)}{n}}] = [\pi_1 ; \pi_2]$$

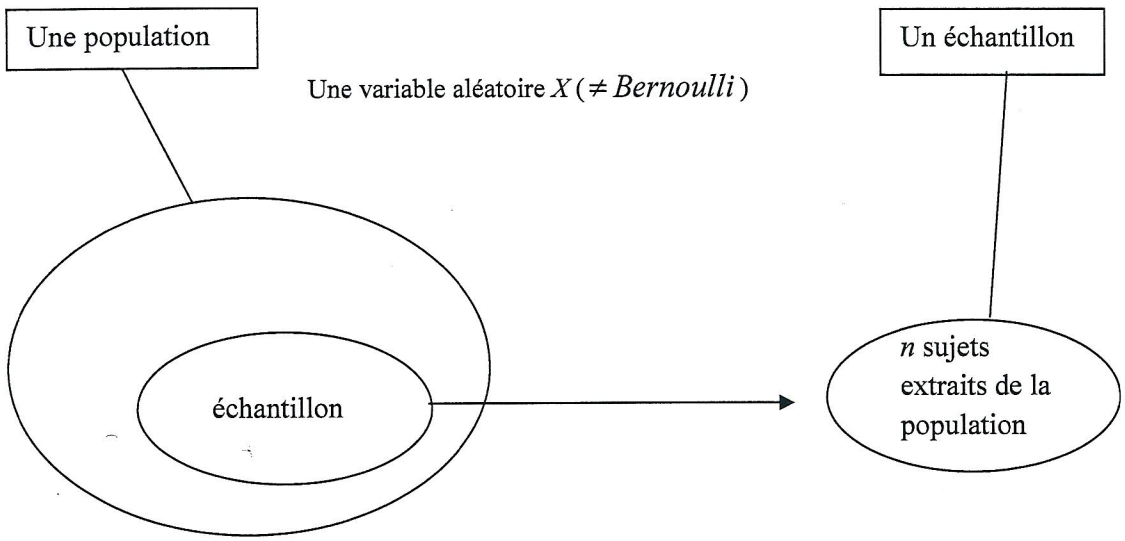
Attention ! Conditions de validité à vérifier **aux bornes** de l'intervalle, après son calcul :

$$n\pi_1 \geq 5 \quad n(1-\pi_1) \geq 5 \quad n\pi_2 \geq 5 \quad n(1-\pi_2) \geq 5$$

Précision de l'IC : $\varepsilon_\alpha \sqrt{\frac{p(1-p)}{n}}$ largeur de l'IC : $2 \times \varepsilon_\alpha \sqrt{\frac{p(1-p)}{n}}$

Remarque : Des échantillons différents fournissent des intervalles de confiance différents.

MOYENNE : Ne pas confondre Intervalle de Pari et Intervalle de Confiance



La moyenne μ_X de X dans la population (dite vraie ou théorique) est une constante	La moyenne observée m de n valeurs de X est variable d'un échantillon à l'autre (m est une réalisation de M_n)

Pari

à condition que : X Normale ou $n \geq 30$ (TCL)

μ_X est connue σ_X est connu

$$IP_{niv\,1-\alpha}(M_n) = [\mu(M_n) \pm \varepsilon_\alpha \sigma(M_n))] = [\mu_X \pm \varepsilon_\alpha \frac{\sigma_X}{\sqrt{n}}]$$

Quelle moyenne m observée de n valeurs ?

μ_X est inconnue σ_X est inconnu

la moyenne observée m et l'écart-type observé s de n valeurs sont connus

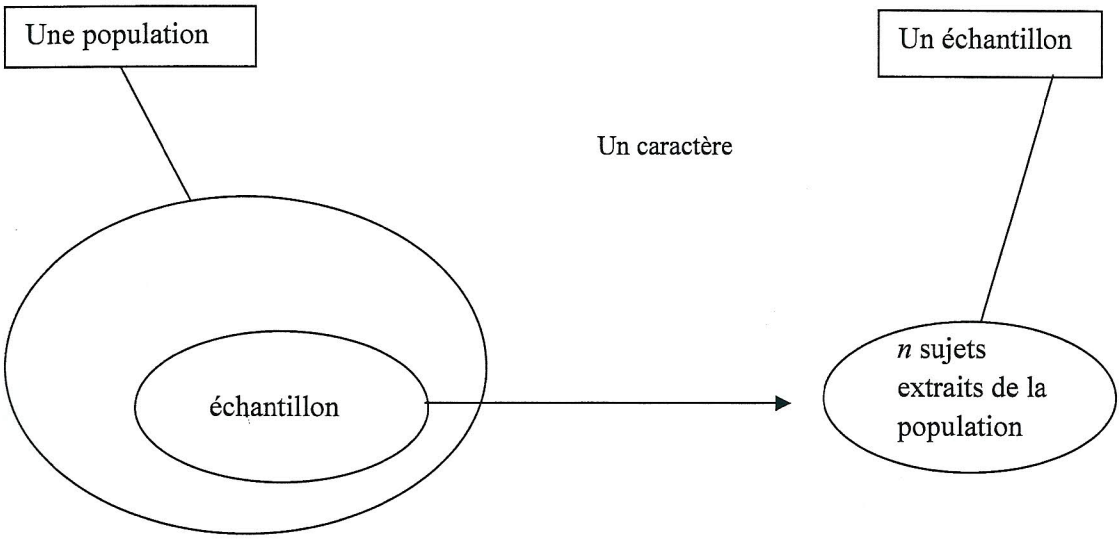
à condition que X Normale ou $n \geq 30$ (TCL)

$$IC_{niv\,1-\alpha}(\mu) = [m \pm \varepsilon_\alpha \frac{s}{\sqrt{n}}]$$

Quelle moyenne vraie dans la population ?

Confiance.

PROPORTION : Ne pas confondre Intervalle de Pari et Intervalle de Confiance



La proportion π de sujets de la population (dite vraie ou théorique) possédant un certain caractère est une constante	La proportion observée p de sujets possédant ce caractère parmi n sujets est variable d'un échantillon à l'autre (p est une réalisation de P_n)

à condition que : $n\pi \geq 5$ *et* $n(1-\pi) \geq 5$

π est connue

$$IP_{niv\,1-\alpha}(P_n) = [\pi \pm \varepsilon_\alpha \sqrt{\frac{\pi(1-\pi)}{n}}]$$

Quelle proportion observée p parmi n sujets ?

π est inconnue

La proportion observée p de sujets possédant ce caractère parmi n sujets est connue

$$IC_{niv\,1-\alpha}(\pi) = [p \pm \varepsilon_\alpha \sqrt{\frac{p(1-p)}{n}}] = [\pi_1 ; \pi_2]$$

Quelle proportion vraie dans la population ?

valide à condition qu'après **calcul** les bornes vérifient :

$n\pi_1 \geq 5 \quad n(1-\pi_1) \geq 5 \quad n\pi_2 \geq 5 \quad n(1-\pi_2) \geq 5$

Test de comparaison de deux moyennes sur échantillons APPARIÉS

Appariés= Non Indépendants

- Un seul groupe de sujets d'effectif n
- Condition de validité : $n \geq 30$
- Deux variables X_A *et* X_B prenant deux valeurs pour chaque sujet du groupe

On effectue pour la variable aléatoire différence $Y = X_A - X_B$

le test de comparaison d'une moyenne observée à la valeur zéro

(test de comparaison d'une moyenne à une norme, à une valeur donnée)

$H_0 : \mu = 0$ $\quad$ $H_1 : \mu \neq 0$

On calcule $z = \dfrac{m}{\sqrt{\dfrac{s^2}{n}}}$

Avec un risque de 1ère espèce fixé à α:

Si $|z| > \varepsilon_\alpha$ alors **RH$_0$** au risque α , et on dit que la moyenne observée des différences est *significativement* différente de zéro et on conclut : les moyennes vraies μ_A *et* μ_B sont différentes.

Degré de signification=
$P(|Z| > |z_{calculé}|)$

sinon **NRH$_0$** et on dit que la moyenne observée des différences n'est pas *significativement* différente de zéro,

mais attention, on ne conclut pas H$_0$, on ne conclut pas que les moyennes vraies μ_A *et* μ_B sont égales !

Pourquoi une procédure différente pour comparer deux moyennes quand les échantillons ne sont pas indépendants ?

X_A *et* X_B ne sont pas indépendantes,

donc M_A (variable Moyenne de n répétitions de X_A) et M_B (variable Moyenne de n répétitions de X_B) ne sont pas indépendantes et dans ce cas :

$Var(M_A - M_B) = Var(M_A) + Var(M_B) - 2\,\mathrm{cov}(M_A, M_B)$

$E(M_A - M_B) = \mu_A - \mu_B$.

La différence entre les deux valeurs relevées est calculée pour chaque sujet : on dispose de n réalisations de la variable différence $Y = X_A - X_B$ dont la moyenne **vraie** est $\mu = \mu_A - \mu_B$

m *et* s^2 sont la moyenne observée et la variance observée de ces différences

Tests du chi-deux

Test de comparaison d'une distribution observée à une distribution théorique : χ^2 d'adéquation

Soit une variable qualitative à k modalités.
On observe sur un échantillon de taille n l'effectif de chaque modalité : les effectifs observés sont $O_1, O_2, \ldots, O_k$

H_0 : la distribution vraie de la variable est égale à la distribution théorique de référence.

H_1 : la distribution de la variable diffère de la distribution théorique de référence.

On calcule les effectifs attendus **sous H_0** : $C_1, C_2, \ldots, C_k$

Condition de validité : tous les $C_i \geq 5$

Sous H_0, $K = \sum_{i=1}^{k} \frac{(O_i - C_i)^2}{C_i} \sim \chi^2 à\ (k-1) ddl$

Avec $\alpha = 5\%$

Si K> $\chi^2_{0,05}(k-1)$ (valeur lue dans la table du χ^2) alors RH_0

Degré de signification : $P(\chi^2(k-1) > K\ calculé)$

Sinon NRH_0

Dans le cas d'une variable à 2 modalités (k=2, nombre de ddl=1)
test équivalent au test de comparaison d'une proportion observée à une proportion théorique, même conclusion.
$K = z^2$

Test de comparaison de plusieurs distributions observées : χ^2 d'homogénéité

Test d'indépendance entre variables qualitatives : χ^2 d'indépendance

Il s'agit de comparer les répartitions d'une variable qualitative à k modalités, observées dans m échantillons.
H_0 : les distributions vraies de la variable sont les mêmes dans les m populations.
H_1 : les distributions vraies sont différentes.

Ou bien il s'agit de tester l'indépendance de deux variables qualitatives, l'une à k modalités, l'autre à m modalités observées dans un échantillon
H_0 : les deux variables sont indépendantes.
H_1 : les deux variables sont liées.

	x_1	...	x_j	...	x_k	Total ligne
y_1						l_1
.						
.						
y_i			$O_{i,j}$			l_i
.						
.						
.						
y_m						l_m
Total colonne	c_1		c_j			n

On calcule les effectifs attendus **sous H_0** :

Pour chaque case, $C_{i,j} = \frac{l_i \times c_j}{n}$

Condition de validité :
Tous les $C_{i,j} \geq 5$

Sous H_0, $K = \sum_{i,j} \frac{(O_{i,j} - C_{i,j})^2}{C_{i,j}} \sim \chi^2 à\ (k-1)(m-1) ddl$

Avec $\alpha = 5\%$

Si K> $\chi^2_{0,05}(k-1)(m-1)$ (valeur lue dans la table du χ^2) alors RH_0

Degré de signification : $P(\chi^2(k-1)(m-1) > K\ calculé)$

Sinon NRH_0

Dans le cas de variables à 2 modalités (k=m=2, nombre de ddl=1)
Test équivalent au test de comparaison de deux proportions observées, même conclusion.
$K = z^2$

Tests d'hypothèses : généralités

On teste une hypothèse H_0 (hypothèse nulle) contre une hypothèse H_1 (hypothèse alternative).

Une règle de décision étant fixée, soit on rejette H_0 (RH_0), soit on ne rejette pas H_0 (NRH_0).

Erreurs et risques d'erreur

Erreur de **1ère espèce** (de type I)=erreur de décision sous H_0 (si H_0 est vraie) =RH_0/H_0
ne peut se produire que lorsqu'on rejette H_0

α=Risque de **1ère espèce**=probabilité de commettre l'erreur de 1ère espèce= $P(RH_0/H_0)$

Erreur de **2nde espèce** (de type II)=erreur de décision sous H_1 (si H_1 est vraie)= NRH_0/H_1
ne peut se produire que lorsqu'on ne rejette pas H_0

β=Risque de **2nde espèce**=probabilité de commettre l'erreur de 2nde espèce= $P(NRH_0/H_1)$

Erreur totale = (RH_0 et H_0 vraie) ou (NRH_0 et H_1 vraie)
$P(\text{erreur totale})=P(RH_0 \cap H_0) + P(NRH_0 \cap H_1) = P(RH_0/H_0)P(H_0) + P(NRH_0/H_1)P(H_1)$

Puissance

Puissance= $P(RH_0/H_1) = 1 - P(NRH_0/H_1) = 1 - \beta$

Degré de signification

Ne se calcule que si on rejette H_0

En cas de rejet de H_0, le degré de signification, noté p, est la plus petite valeur de risque α avec lequel on aurait encore rejeté H_0.

Le degré de signification mesure la force avec laquelle on rejette H_0 : plus il est petit, plus confortable est le rejet.

Tests usuels sur des proportions avec des grands échantillons			
Test de comparaison d'une proportion observée à une proportion théorique donnée π_0	**Un seul** échantillon d'effectif n	**Conditions de validité :** $n\pi_0 \geq 5$ *et* $n(1-\pi_0) \geq 5$	$H_0 : \pi = \pi_0 \qquad H_1 : \pi \neq \pi_0$ Sous H_0, $P_n \sim N(\pi_0 ; \frac{\pi_0(1-\pi_0)}{n})$ donc $Z = \frac{P_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(0;1)$ On calcule $z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$ avec p la proportion observée dans l'échantillon Avec un risque de 1ère espèce fixé à α: Si $\lvert z \rvert > \varepsilon_\alpha$ alors **RH_0** au risque α , et on dit que la proportion observée est *significativement* différente de la valeur donnée et on conclut : la proportion vraie π est différente de la valeur donnée π_0. **Degré de signification**= $P(\lvert Z \rvert > \lvert z_{calculé} \rvert)$ sinon **NRH_0** et on dit que la proportion observée n'est pas *significativement* différente de la valeur donnée, mais attention, on ne conclut pas H_0, on ne conclut pas que la proportion vraie est égale à la valeur donnée π_0.
Test de comparaison de deux proportions observées	**Deux** échantillons d'effectifs n_A et n_B	**Conditions de validité :** on calcule d'abord la proportion observée p **commune** dans les deux échantillons $p = \frac{n_A p_A + n_B p_B}{n_A + n_B}$ avec p_A et p_B les proportions observées dans chaque groupe puis on vérifie que : $n_A p \geq 5$ *et* $n_A(1-p) \geq 5$ $n_B p \geq 5$ *et* $n_B(1-p) \geq 5$	$H_0 : \pi_A = \pi_B (= \pi) \qquad H_1 : \pi_A \neq \pi_B$ Sous H_0, $P_{n_A} - P_{n_B} \sim N(0; \frac{\pi(1-\pi)}{n_A} + \frac{\pi(1-\pi)}{n_B} = \pi(1-\pi)(\frac{1}{n_A} + \frac{1}{n_B}))$ Donc $Z = \frac{P_{n_A} - P_{n_B}}{\sqrt{\pi(1-\pi)(\frac{1}{n_A} + \frac{1}{n_B})}} \sim N(0;1)$ On calcule $z = \frac{p_A - p_B}{\sqrt{p(1-p)(\frac{1}{n_A} + \frac{1}{n_B})}}$ Avec un risque de 1ère espèce fixé à α: Si $\lvert z \rvert > \varepsilon_\alpha$ alors **RH_0** au risque α , et on dit que les proportions observées sont *significativement* différentes et on conclut : les proportions vraies π_A et π_B sont différentes. **Degré de signification**= $P(\lvert Z \rvert > \lvert z_{calculé} \rvert)$ sinon **NRH_0** et on dit que les proportions observées ne sont pas *significativement* différentes, mais attention, on ne conclut pas H_0, on ne conclut pas que les proportions vraies π_A et π_B sont égales.

<table>
<tr><th colspan="4">Tests usuels sur des moyennes avec des grands échantillons</th></tr>
<tr>
<td>Test de comparaison d'une moyenne observée à une valeur donnée</td>
<td>Un seul échantillon d'effectif n</td>
<td>Conditions de validité :

$n \geq 30$</td>
<td>$H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0$
Sous H_0 , $M_n \sim N(\mu_0 ; \frac{\sigma^2}{n})$ donc $Z = \frac{M_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0;1)$ On calcule $z = \frac{m - \mu_0}{\sqrt{\frac{s^2}{n}}}$
avec m la moyenne observée dans l'échantillon et s^2 la variance observée dans l'échantillon.
Avec un risque de 1ère espèce fixé à α: Si $|z| > \varepsilon_\alpha$ alors RH$_0$ au risque α , et on dit que la moyenne observée est significativement différente de la valeur donnée et on conclut : la moyenne vraie μ est différente de la valeur donnée μ_0.
Degré de signification= $P(|Z| > |z_{calculé}|)$

sinon NRH$_0$ et on dit que la moyenne observée n'est pas significativement différente de la valeur donnée, mais attention, on ne conclut pas H_0, on ne conclut pas que la moyenne vraie est égale à la valeur donnée μ_0.</td>
</tr>
<tr>
<td>Test de comparaison de deux moyennes observées</td>
<td>Deux échantillons d'effectifs n_A et n_B</td>
<td>Conditions de validité :

$n_A \geq 30$
et
$n_B \geq 30$</td>
<td>$H_0 : \mu_A = \mu_B \qquad H_1 : \mu_A \neq \mu_B$
Sous H_0 , $M_{n_A} - M_{n_B} \sim N(0; \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B})$ donc $Z = \frac{M_{n_A} - M_{n_B}}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \sim N(0;1)$
On calcule $z = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$ avec m_A et m_B les moyennes observées dans les échantillons et $s_A^{\,2}$ et $s_B^{\,2}$ les variances observées dans les échantillons.
Avec un risque de 1ère espèce fixé à α: Si $|z| > \varepsilon_\alpha$ alors RH$_0$ au risque α , et on dit que les moyennes observées sont significativement différentes et on conclut : les moyennes vraies μ_A et μ_B sont différentes. Degré de signification= $P(|Z| > |z_{calculé}|)$

sinon NRH$_0$ et on dit que les moyennes observées ne sont pas significativement différentes, mais attention, on ne conclut pas H_0, on ne conclut pas que les moyennes vraies μ_A et μ_B sont égales.</td>
</tr>
</table>

Calcul de la puissance

Test de comparaison de deux proportions observées

$H_0 : \pi_A = \pi_B (= \pi)$ $\quad H_1 : \pi_A - \pi_B = \Delta$

Sous H_1, $Z' = \dfrac{P_{n_A} - P_{n_B}}{\sqrt{\pi(1-\pi)(\frac{1}{n_A}+\frac{1}{n_B})}} \sim N(\dfrac{\Delta}{\sqrt{p(1-p)(\frac{1}{n_A}+\frac{1}{n_B})}};1)$

$\alpha = 5\%$

Puissance = $P(RH_0/H_1)$

$= P(|Z'| > 1{,}96)$

$= P(Z' < -1{,}96) \;+\; P(Z' > 1{,}96)$

$= P(Z < -1{,}96 - \dfrac{\Delta}{\sqrt{p(1-p)(\frac{1}{n_A}+\frac{1}{n_B})}}) \;+\; P(Z > 1{,}96 - \dfrac{\Delta}{\sqrt{p(1-p)(\frac{1}{n_A}+\frac{1}{n_B})}})$

L'un des deux termes est en général négligeable.

Test de comparaison de deux moyennes observées

$H_0 : \mu_A = \mu_B$ $\quad H_1 : \mu_A - \mu_B = \Delta$

Sous H_1, $Z' = \dfrac{M_{n_A} - M_{n_B}}{\sqrt{\frac{\sigma_A^2}{n_A}+\frac{\sigma_B^2}{n_B}}} \sim N(\dfrac{\Delta}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}};1)$

$\alpha = 5\%$

Puissance = $P(RH_0/H_1)$

$= P(|Z'| > 1{,}96)$

$= P(Z' < -1{,}96) \;+\; P(Z' > 1{,}96)$

$= P(Z < -1{,}96 - \dfrac{\Delta}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}}) \;+\; P(Z > 1{,}96 - \dfrac{\Delta}{\sqrt{\frac{s_A^2}{n_A}+\frac{s_B^2}{n_B}}})$

L'un des deux termes est en général négligeable.

Test de comparaison d'une proportion observée à une proportion théorique donnée π_0

$H_0 : \pi = \pi_0$ $\quad H_1 : \pi - \pi_0 = \Delta$

Sous H_1, $Z' = \dfrac{P_n - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \sim N(\dfrac{\Delta}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}};1)$

$\alpha = 5\%$

Puissance = $P(RH_0/H_1)$

$= P(|Z'| > 1{,}96)$

$= P(Z' < -1{,}96) \;+\; P(Z' > 1{,}96)$

$= P(Z < -1{,}96 - \dfrac{\Delta}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}) \;+\; P(Z > 1{,}96 - \dfrac{\Delta}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}})$

L'un des deux termes est en général négligeable.

Test de comparaison d'une moyenne observée à une valeur donnée

$H_0 : \mu = \mu_0$ $\quad H_1 : \mu - \mu_0 = \Delta$

Sous H_1, $Z' = \dfrac{M_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(\dfrac{\Delta}{\sqrt{\frac{s^2}{n}}};1)$

$\alpha = 5\%$

Puissance = $P(RH_0/H_1)$

$= P(|Z'| > 1{,}96)$

$= P(Z' < -1{,}96) \;+\; P(Z' > 1{,}96)$

$= P(Z < -1{,}96 - \dfrac{\Delta}{\sqrt{\frac{s^2}{n}}}) \;+\; P(Z > 1{,}96 - \dfrac{\Delta}{\sqrt{\frac{s^2}{n}}})$

L'un des deux termes est en général négligeable.