1 Bases

- Ω est l'univers (les valeurs possibles)
- $X : \Omega \to \mathbb{N}$ (ex: $\Omega \mapsto 1, 2, 3, 4, 5, 6$) pour les dés
  - Discrète: énumération des éléments
  - Continue: tu connais
- $X = 2$ est un évènement. $X = 2 \vee X = 3$ une liste d'év.
- $P(X = 2)$ la probabilité de l'évènement.
  - $\forall E \subseteq \Omega, P(E) \in [0, 1]$ et $P(\Omega) = 1$;
  - $\forall E \subseteq \Omega, P(E) = 1 - P(E^c)$, où E^c est l'événement complémentaire de E;
  - $\forall E_1, E_2 \subseteq \Omega, P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$;
  - $\forall E_1, E_2 \subseteq \Omega, E_1 \subseteq E_2 \Rightarrow P(E_1) \leq P(E_2)$.

1.1 Fonction de répartition

$F_X : \mathbb{R} \to [0, 1], \quad x \mapsto P(X \leq x)$.

- Si discrète:
  1. $P_X(X = x_i) = F_X(x_i) - F_X(x_{i-1})$
  2. $F_X(x) = \sum_{x_i \leq x} p_X(x_i)$
- Si continue:
  1. $P(X \leq a) = P(X < a)$
  2. $P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(u)\,du$

1.2 Quantile

Définition : Quantile d'ordre q d'une variable aléatoire X: x_q tel que $P(X \leq x_q) = q$ ou $F_X(x_q) = q$.

Exemple : La médiane est le quantile lorsque $q = 0.5$.

1.3 Espérance

- $X = g(X)$
- $E(Y) = \sum_i g(x_i) p_X(x_i)$ dans le cas discret.
- $E(Y) = \int_{-\infty}^{+\infty} g(x) f_X(x) dx$ dans le cas continu.

L'espérance est linéaire:

- $E(X + Y) = E(X) + E(Y)$
- $E(aX + b) = aE(X) + b$

Soit X une variable aléatoire, et $Y = ag(X) + bh(X)$, où a et b sont deux constantes et g et h deux fonctions. Alors on a : $E(Y) = aE(g(X)) + bE(h(X))$.

1.4 Variance

$V(X) = E((X - \mu)^2) = \sigma^2$ et $\sigma = \sqrt{V(X)}$.

- Décentrage de la variance: $\mu_2 = E(X^2) - \mu^2$.
- $V(aX + b) = a^2 V(X)$.

2 Les lois du cours

2.1 Loi uniforme discrête

$P(X = x_i) = \frac{1}{r}$, pour $x_i \in \{1, \ldots, r\}$

- $E(X) = \frac{r+1}{2}$
- $V(X) = \frac{r^2-1}{12}$

2.2 Loi de Bernoulli $\mathcal{B}(p)$

$P(X = 1) = p$, $P(X = 0) = 1 - p$

2.3 Loi binomial $\mathcal{B}(n, p)$

$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$

- $E(X) = np$
- $V(X) = np(1 - p)$

Si $X \sim B(n_1, p)$ et $Y \sim B(n_2, p)$, alors $X + Y \sim B(n_1 + n_2, p)$.

2.4 Loi uniforme continue $\mathcal{U}([a, b])$

$$p(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{sinon} \end{cases}$$

$$F(x) = \begin{cases} 0 & \text{si } x < a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{sinon} \end{cases}$$

Si $X \sim U([a, b])$, alors

- $E(X) = \frac{a+b}{2}$
- $V(X) = \frac{(b-a)^2}{12}$

2.5 Loi de Gauss $\mathcal{N}(\mu, \sigma^2)$

$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

La loi normale standardisée, $\mathcal{N}(0, 1)$, est utilisée pour simplifier la représentation des variables normales : $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

- $E(X) = \mu$
- $V(X) = \sigma^2$

2.6 Moyenne échantillon/empirique

- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- $E(\bar{X}) = \mu$
- $V(\bar{X}) = \frac{\sigma^2}{\sqrt{n}}$
- $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ (variance de l'échantillon)
- $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$ où N est la taille de la population (variance de la population)

2.7 Théorème central limite

Définition : Soit $\overline{X_n}$, la moyenne empirique d'un n-échantillon aléatoire de loi mère quelconque, de moyenne théorique μ et de variance théorique σ^2. Alors, pour n ($n \geq 30$) assez grand, $\overline{X_n}$ suit approximativement une loi $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

2.8 Application du Théorème Central Limite:

Exemple : Dans une usine de bonbons, la cadence des salariés étudie les bonbons en moyenne avec un écart type de 3.

Loi du nombre de bonbons par sachet ?

On tire 30 sachets et on fait la moyenne du nombre de bonbons.

Quelle est la probabilité d'avoir entre 99 et 101 en moyenne?

X_i: nombre de bonbons dans le sachet i.

- $\overline{X} = \frac{1}{30} \sum_{i=1}^{30} X_i$ par le TCL, $\overline{X} - \mu \sim \mathcal{N}(0, 1)$ équivaut.

3 Statistique de test

Définition : Dans le cas d'hypothèses simples, l'espace paramétrique est $\Theta = \{\theta_0, \theta_1\}$, et le test vise à décider entre $H_0 : \theta = \theta_0$ et $H_1 : \theta = \theta_1$. Un test pour H_0 est une règle de décision fondée sur la valeur réalisée t sur un échantillon, d'une statistique T, appelée *statistique du test*, à valeurs dans $\mathbb{R}$.

3.1 Région d'acceptation

La règle de décision pour le test est la suivante:

- Si $T \in A$ (une partie de $\mathbb{R}$), on accepte H_0.
- Si $T \notin A$ (le complémentaire de A), on rejette H_0.

A est appelée *région d'acceptation* du test, et $\bar{A}$ la *région de rejet* du test. Si A est un intervalle, il est appelé *intervalle de confiance* lié au test.

4 Risques

4.1 Risque de première espèce

Le *risque de première espèce*, noté α, est la probabilité de rejeter H_0 alors qu'elle est vraie: $\alpha = P_{H_0}(T \in \bar{A})$.

4.2 Risque de deuxième espèce

Le *risque de deuxième espèce*, noté β, est la probabilité d'accepter H_0 alors qu'elle est fausse: $\beta = P_{H_1}(T \in A)$.

Les risques α et β sont contrôlables et dépendent de la taille de l'échantillon et du choix de la région d'acceptation A.

5 Construction d'un test

Pour construire un test, on détermine les hypothèses H_0 et H_1, on recherche une statistique pertinente T, on fixe un niveau α, on détermine l'intervalle de confiance associé à ce niveau, et on prend une décision en fonction de la réalisation de T sur l'échantillon.

6 Puissance d'un test

La *puissance d'un test* est la probabilité de rejeter H_0 lorsque H_1 est vraie: $P(T \in \bar{A} | H_1)$. Un test est *sans biais* si sa puissance est supérieure ou égale à son niveau α.

7 Exercices

Exercice Limite centrale:

Soit $X_1, \ldots, X_n$ n variables aléatoires normales $\mathcal{N}(m, \sigma^2)$ et indépendantes. On note $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$ la moyenne empirique. Quelle est la loi de $\overline{X}$?

Le théorème de la limite centrale généralise le résultat précédent à une variable aléatoire X quelconque. Si $X_1, \ldots, X_n$ est une suite de variables aléatoires de même loi et indépendantes, ayant une moyenne théorique m et un écart-type σ, alors la moyenne empirique $\overline{X} = \frac{1}{n} \sum_{i=1}^n X_i$ suit approximativement la loi normale $\mathcal{N}(m, \frac{\sigma^2}{n})$.

1. La valeur attendue de $\overline{X}$, $E(\overline{X})$, se calcule comme suit : $E(\overline{X}) = E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \times n \times m = m$ car la valeur attendue de chaque X_i est m.
2. La variance de $\overline{X}$, $V(\overline{X})$, se calcule comme suit : $V(\overline{X}) = V\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(X_i) = \frac{1}{n^2} \times n \times \sigma^2 = \frac{\sigma^2}{n}$ car la variance de chaque X_i est σ^2 et les X_i sont indépendants.
3. En conclusion, la moyenne empirique $\overline{X}$ suit une loi normale $\mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$.

Exercice Application à Bernouilli:
Appliquer ce théorème au cas de n variables indépendantes de Bernoulli $Be(p)$. Que vaut $\overline{X}$? En déduire une approximation de la loi binomiale par une loi normale.

1. La loi de Bernoulli $Be(p)$ a une moyenne théorique $m = p$ et une variance $\sigma^2 = p(1-p)$.
2. La moyenne empirique est donnée par $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Selon le théorème de la limite centrale, pour un grand nombre d'essais n, $\overline{X}$ suit approximativement une loi normale. La moyenne de cette distribution normale est la même que celle de la loi binomiale, soit p, et la variance est $\frac{p(1-p)}{n}$.
3. Pour un cas spécifique, comme le lancer d'une pièce équilibrée ($p = 0.5$) avec $n = 100$ lancers, la moyenne empirique $\overline{X}$ suit approximativement la loi normale $\mathcal{N}(50, 0.25)$. Ici, 50 est la moyenne (car $np = 100 \times 0.5$) et 0.25 est la variance (car $\frac{0.5\times(1-0.5)}{100} = 0.25$). L'écart-type est alors $\sqrt{0.25} = 0.5$.

A) Modélisation de densités de probabilité (6 pts)
Soit les 3 distributions $H1(x)$, $H2(x)$ et $H3(x)$, toutes composées de 100 événements, représentant les occurrences pour x compris entre 0 et 31 : $H1(x) = \{...\}$; $H2(x) = \{...\}$; $H3(x) = \{...\}$;

1. À partir de ces 3 distributions, calculer et tracer les densités de probabilité (ddp) correspondantes $f1(x)$, $f2(x)$ et $f3(x)$.
2. À partir de la fonction générique $f(x) = C\exp(-K|x-\mu|^{\alpha})$ en déduire la valeur de α la plus pertinente pour chacune des ddp $f1(x)$, $f2(x)$ et $f3(x)$. Pour rappel, $\alpha = 0$: distribution uniforme; $\alpha = 1$: distribution Laplacienne; $\alpha = 2$: distribution Gaussienne.
3. Approcher la ddp $f2(x)$ par une distribution Gaussienne en calculant $moy2$ et $sigm2$, correspondant à la valeur moyenne et l'écart type (donner la formule obtenue). Tracer cette distribution sur la courbe représentant $f2(x)$.
4. Approcher la ddp $f3(x)$ par une distribution Gaussienne en calculant $moy3$ et $sigm3$, correspondant à la valeur moyenne et l'écart type (donner la formule obtenue). Tracer cette distribution sur la courbe représentant $f3(x)$.
5. Qu'en déduisez-vous ? Comment mesurer la distance entre une ddp et la distribution Gaussienne qui l'approche ?

1. On divise ici chaque valeur par le nombre de pixel total. $\forall x, f(x) = H1(x)/100$ puis on les mets dans un graphe.
2. - **Laplacienne :** Elle est caractérisée par sa forme en cloche avec des queues lourdes.
   - **Gaussienne (ou Normale) :** Cloche symétrique autour de sa moyenne.
3. La moyenne est calculée par la formule : $moy2 = (\sum_{i=0}^{31} x_i \times f2(x_i))/$ où x_i sont les valeurs de x (de 0 à 31) et $f2(x_i)$ sont les probabilités associées à ces valeurs.

   L'écart type est calculé en utilisant la formule : $sigm2 = \sqrt{\sum_{i=0}^{31}(x_i - moy2)^2 \times f2(x_i)}$
4. Pareil
5. (a) **Erreur Quadratique Moyenne (EQM)**: $EQM = \frac{1}{n}\sum_{i=0}^{n-1}(f(x_i) - G(x_i))^2$ où $f(x_i)$ est la valeur de la ddp observée et $G(x_i)$ est la valeur de la distribution gaussienne pour chaque x_i.

Mélange de 2 Gaussiennes
Soit la densité de probabilité (ddp) suivante $f(x)$ représentant les probabilités pour x compris entre 0 et 31 : $H(x) = \{...\}$

1. Tracer $f(x)$.
2. En considérant que $f(x)$ est un mélange de 2 gaussiennes, indiquer la valeur des 2 modes (valeurs maximales) et proposer une valeur de seuil séparant les 2 modes.
3. Pour chacun des modes, calculer la valeur moyenne et l'écart type μ_1, σ_1, μ_2, σ_2.
4. Par rapport aux probabilités par mode (β_1 et β_2), proposer un modèle de mélange de 2 Gaussiennes - paramètres à introduire μ_1, σ_1, μ_2, σ_2 et β_1 (sachant que $\beta_2 = 1 - \beta_1$).

1. Easy
2. On prend au niveau du "creux". Dans ce cas c'était $x = 6$
3. **Calculs pour le premier mode:**

   La moyenne μ_1 est la moyenne pondérée des valeurs de x:

   $\mu_1 = \frac{\sum_{i=0}^{6} x_i \times H(x_i)}{\sum_{i=0}^{6} H(x_i)}$

   L'écart type σ_1 est la racine carrée de la moyenne pondérée des carrés des écarts de chaque valeur de x par rapport à μ_1:

   $\sigma_1 = \sqrt{\frac{\sum_{i=0}^{6}(x_i-\mu_1)^2 \times H(x_i)}{\sum_{i=0}^{6} H(x_i)}}$

   Calculs pour le second mode:

   La moyenne μ_2 est calculée de façon similaire au premier mode:

   $\mu_2 = \frac{\sum_{i=7}^{31} x_i \times H(x_i)}{\sum_{i=7}^{31} H(x_i)}$

   L'écart type σ_2 est également calculé de façon similaire:

   $\sigma_2 = \sqrt{\frac{\sum_{i=7}^{31}(x_i-\mu_2)^2 \times H(x_i)}{\sum_{i=7}^{31} H(x_i)}}$

   En effectuant ces calculs, nous obtenons les résultats suivants pour les deux modes :
   - Pour le premier mode :
     – Moyenne (notée μ_1) : 4.47
     – Écart type (noté σ_1) : 1.09
   - Pour le second mode :
     – Moyenne (notée μ_2) : 11.00
     – Écart type (noté σ_2) : 2.17