MOSE2014 : Probabilités et Statistiques

Philippe Thieullen
Institut de Mathématiques
Université Bordeaux 1, CNRS, UMR 5251
F-33405 Talence, France
Philippe.Thieullen@math.u-bordeaux1.fr

Talence, 10 janvier 2012

Résumé du programme

A prévoir :
- 18 séances de 1h20 : cours-td intégrés.
- 4 séance de 1h20 de TP machine à coefficient 0.15 : travail sur le logiciel R ; les étudiants peuvent s'associer par binôme ou trinôme et envoient par courrier électronique leur fichier source en fin de séance.
- 3 contrôles continus de 30 mn à coefficient 0.15 : chaque groupe organise son propre contrôle tout en respectant la cadence des séances.
- 1 DS de 1h30 à coefficient de 0.3 : le DS est commun à l'ensemble de l'UE mais chaque enseignant corrige son propre groupe. Il faut donc prévoir de réunir les copies le jour de l'examen par groupe.
- 1 DST de 1h30 à coefficient de 0.4 : mêmes conditions que pour le DS.

1. **Statistique descriptive et Indicateurs numériques**
   - Terminologie (population, échantillon $(x_1, x_2, \cdots, x_n)$, taille, caractères, modalités).
   - Notion de caractère statistique (quantitatif, qualitatif, discret, continu), classe (amplitude, milieu).
   - Représentation des données d'un seul caractère : série brute, tableau par valeurs-effectifs (ξ_i, n_i) et par classes-effectifs $([\xi_{i-1}, \xi_i[, n_i)$, par fréquence-effectifs, $f_i = n_i/n$.
   - Diagramme en bâton pour des variables qualitatives, histogramme pour des variables numériques continues (discuter en exercices le cas des classes n'ayant pas toutes la même amplitude), courbe des effectifs cumulés.
   - Moyenne observée : cas d'une série brute, cas d'un échantillon donné par valeurs-effectifs,

$$\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}(n_1\xi_1 + \cdots + n_r\xi_r)$$

– Médiane $m = q_{50\%}$. Définition dans le cas d'une série brute réordonnée. Formule graphique utilisant la courbe des effectifs cumulés $N_1, \dots, N_r$ ou des fréquences cumulées, $(F_1, \cdots, F_r)$, dans le cas continu. Formule théorique de la médiane pour un échantillon donné par classes-effectifs

$$\frac{q_{50\%} - \xi_{i-1}}{\xi_i - \xi_{i-1}} = \frac{50\% - F_{i-1}}{F_i - F_{i-1}} = \frac{\frac{1}{2}n - N_{i-1}}{N_i - N_{i-1}}.$$

– Variance s^2_{n-1} ou écart-type observée s_{n-1},

$$s^2_{n-1} = \frac{1}{n-1}\Big((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\Big).$$

– Premier quartile $q_{25\%}$, troisième quartile $q_{75\%}$, intervalle interquartile, boxplot (de préférence à boîte à moustache) comme mesure de la dispersion, dans le cas d'une représentation de données par classes-effectifs.
– En TP machine, on verra comment déterminer les quartiles dans le cas d'une série brute de taille n réordonnées par ordre croissant,

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)},$$
$$q_{25\%} = x_{(n/4)}, \quad m = q_{50\%} = x_{(n/2)}, \quad q_{75\%} = x_{(3n/4)}.$$

2. **Espace et mesure de probabilité**
   – Notions d'espace fondamental Ω (ou ensemble des épreuves), d'événements $A \subset \Omega$, d'événements élémentaires $\omega \in \Omega$.
   – Utiliser des exemples concrets. Construire explicitement $(\Omega, \mathbb{P})$ dans chaque cas (ne pas introduire d'algèbre d'événements !).
   – Opérations sur les événements : événement mutuellement incompatibles (ou disjoints $A \cap B = \emptyset$), événement contraire $\bar{A}$ (notation commune imposée), événement certain, impossible.
   – Probabilité d'un événement. Probabilité du complémentaire (insister sur son utilisation), de la réunion d'ensembles disjoints (deux ou plusieurs). Formule générale

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

3. **Exercices de révision**
4. **Indépendance et probabilités conditionnelles**
   – Indépendance d'événements : définition, exemples.
   – Notion d'événements conditionnels. Définition de la probabilité conditionnelle de A sachant B, $\mathbb{P}(A \,|\, B)$ (notation à privilégier sur $\mathbb{P}_B(A)$)l.
   – Formule des probabilités composées : $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A)$.
   – Formule des probabiltés totales : système complet d'événements ou partition $\Omega = A_1 \cup \cdots A_r$, formule

$$\mathbb{P}(B) = \mathbb{P}(A_1)\mathbb{P}(B|A_1) + \cdots + \mathbb{P}(A_r)\mathbb{P}(B|A_r).$$

- Formule de Bayes. (Un exemple type : 15% d'individus d'une certaine population présente une affection A. Un test de dépistage est réalisé. Il s'avère que le test donne 95% de résultats positifs pour les personnes atteintes par A et 10% de résultats positifes pour les personnes non atteintes. Une personne prise au hasard subit le test. Si le test est positif, quelle est la probabilité que cette personne soit atteinte par A ? Si le test est négatif, quelle est la probabilité qu'elle soit indemne ?)
- Présenter plutôt Bayes sous forme d'un tableau ou d'une arborescence.

5. **Variables aléatoires discrètes et lois usuelles**
   - Définition générale d'une variable X. Exemple de la loi uniforme, de la loi de Bernoulli. Exemple d'un lancé de dés de la somme des faces de deux dés.
   - Définition de la loi de probabilité, de la fonction de répartitions $F(x) = \mathbb{P}(X \leq x)$. Faire le lien avec la courbe cumulée.
   - Espérance $\mathbb{E}[X]$, variance $\mathrm{Var}(X)$, écart-type σ, espérance d'une fonction de de la variable X, $\mathbb{E}[\phi(X)]$.
   - Espérance, variance d'une somme de v.a. Indépendance de deux v.a.
   - *Loi de Bernoulli* $\mathcal{B}(p)$. Loi d'une variable X prenant deux valeurs $\{0, 1\}$. Ses paramètres sont donnés par

$$\mathbb{P}(X = 1) = p,\ \mathbb{P}(X = 0) = 1 - p, \quad \mathbb{E}[X] = p, \quad \mathrm{Var}(X) = p(1 - p).$$

   - *Loi binomiale* $\mathcal{B}(n, p)$. Loi d'une variable X prenant ses valeurs dans $\{0, 1, \ldots, n\}$. C'est la loi de la somme de n variables indépendantes et de même loi (i.i.d.) égale à une loi de Bernoulli. Ses paramètres sont donnés par

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \mathbb{E}[X] = np, \quad \mathrm{Var}(X) = np(1 - p).$$

   - *Loi de Poisson* $\mathcal{P}(\lambda)$. (Eventuellement en exercice) Loi d'une variable X prenant des valeurs entières quelconques et servant par exemple à modéliser un nombre d'appels téléphoniques par unité de temps. $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$, pour $k = 0, 1, 2, \ldots$, $\mathbb{E}[X] = \lambda$ et $\mathrm{Var}(X) = \lambda$.

6. **Exercices de révision**

7. **Variables aléatoires continues et lois uselles**
   - Définition au moyen de la notion de densité de probabilité $f(x)$. Exemple de la loi uniforme sur $[0, 1]$, sur $[a, b]$.
   - Fonction de répartition $F_X(x) = \mathbb{P}(X \leq x)$ et quantile d'ordre α, $\mathbb{P}(X \leq q_\alpha) = \alpha$ d'une loi à densité.
   - Espérance, variance, écart-type. Calculer explicitement ces trois quantités pour la loi uniforme.
   - Espérance d'une fonction de la variable X, $\mathbb{E}[\phi(X)]$.
   - Cas de plusieurs variables aléatoires. Espérance de la somme, du produit de deux v.a. Cas indépendant : addition des variances.

– *Loi uniforme* $\mathcal{U}(a,b)$. X prend des valeurs dans $[a,b]$ et sa densité est donnée par

$$f(x)=\frac{1}{b-a}\mathbb{1}_{\{a<x<b\}},\quad \mathbb{E}[X]=\frac{b+a}{2},\quad \mathrm{Var}(X)=\frac{(b-a)^2}{12}.$$

– *Loi normale* $\mathcal{N}(\mu,\sigma^2)$ *centrée réduite.* X prend ses valeurs dans $\mathbb{R}$. Ses paramètres sont donnés par

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big),\quad \mathbb{E}[X]=\mu,\quad \mathrm{Var}(X)=\sigma^2.$$

La somme de v.a. normales indépendentes est encore normale.

– *Loi exponentielle* $\mathcal{E}(\theta)$. (Eventuellement en exercice) La variable $X\geq 0$ est positive. Ses paramètres sont donnés par

$$f(x)=\theta^{-1}e^{-\frac{x}{\theta}}\mathbb{1}_{\{x>0\}},\qquad \mathbb{E}[X]=\theta,\quad \mathrm{Var}(X)=\theta^2.$$

– *Loi du chi-deux* $\chi^2(n)$ *à* n *ddl.* X prend ses valeurs dans $\mathbb{R}^+$ et a même loi que la v.a. $Z_1^2+\cdots+Z_n^2$. Ses paramètres sont donnés par (ne pas retenir)

$$f(x)=\frac{x^{n/2-1}e^{-x/2}}{2^{n/2}\Gamma(n/2)},\quad \mathbb{E}[X]=n,\quad \mathrm{Var}(X)=2n.$$

– *Loi de Student* $\mathcal{T}(n)$ *à* n *ddl.* X prend ses valeurs dans $\mathbb{R}$ et a même loi que la v.a. $X=U/\sqrt{V/n}$ où U et V sont indépendantes, U de loi $\mathcal{N}(0,1)$ et V de loi $\chi^2(n)$. Ses paramètres sont donnés par (ne pas retenir)

$$f(x)=\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}\Big(1+\frac{x^2}{n}\Big)^{-\frac{n+1}{2}},\quad \mathbb{E}[X]=0,\quad \mathrm{Var}(X)=\frac{n}{n-2}.$$

– Exemples d'applications des deux prières lois et utilisation des tables numériques (réduction à la loi normale centrée réduite, exemples de calculs).

– **Premier contrôle continu (30 mn)** Contrôle sur l'ensemble des châpitres portant sur les probabilités combinatoires et les variables aléatoires discrètes.

8. **Exercices portant sur les variables aléatoires continues**
9. **Théorème de la limite centrale et applications**

– Epreuves répétées, somme et moyenne

$$Y=X_1+\cdots+X_n,\quad \bar{X}=\frac{1}{n}\Big(X_1+\cdots+X_n\Big).$$

Espérance de $\bar{X}$, variance de $\bar{X}$. Cas de sommes de v.a. indépendentes

$$\mathbb{E}[\bar{X}]=\mathbb{E}[X]=\mu,\quad \mathrm{Var}(\bar{X})=\frac{1}{n}\mathrm{Var}(X_1)=\frac{1}{n}\sigma^2\quad \text{(cas iid)}.$$

Bien comprendre la différence entre $\mathrm{Var}(10X)$ et $\mathrm{Var}(X_1+X_2+\cdots+X_{10})$ pour des v.a. iid. Savoir se ramener à la variable centrée réduite $Z=\sqrt{n}(\bar{X}-\mu)/\sigma$

– Théorème de la limite centrale,

$$\mathbb{P}\Big(x < \sqrt{n}\frac{\bar{X}-\mu}{\sigma} < y\Big) \simeq \int_x^y \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\, dt.$$

ou bien

$$\mathbb{P}\big(n\mu + x\sigma\sqrt{n} < \sum_{i=1}^n X_i < n\mu + y\sigma\sqrt{n}\big) \simeq \int_x^y \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\, dt.$$

– Approximation d'une loi binomiale, loi de $X_1+\ldots+X_n$, lorsque les X_i sont des v.a. indépendantes de même loi $\mathcal{B}(p)$, par la loi normale $\mathcal{N}(np, np(1-p))$ lorsque n est grand,

$$\mathbb{P}\Big(x < \frac{X_1+\ldots+X_n-np}{\sqrt{np(1-p)}} < y\Big) \simeq \int_x^y \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{1}{2}t^2\Big)\, dt.$$

– Utilisation en exercices des tables de ces lois : tables des fonctions de répartition et tables des quantiles.

10. **Estimation ponctuelle et intervalle de confiance I**
   – Statistique inférentielle versus statistique descriptive : réalisation de n v.a. indépendantes $(X_1, \cdots, X_n)$ de loi inconnues $p_\theta(\xi_i)$ dans le cas discret, $p_\theta(x)\, dx$ dans le cas continu
   – Définition d'un estimateur ponctuel d'une quantité $\tau(\theta)$: c'est une v.a. $T(X) = T(X_1, \cdots, X_n)$ fonction uniquement de l'échantillon X, sensée représenter τ. Exemple : $\theta = (\mu, \sigma^2)$ et $\tau(\theta) = \sigma^2$ pour une famille de lois normales $\mathcal{N}(\mu, \sigma^2)$.
   – Qualité d'un estimateur ponctuel : avec ou sans biais

$$\mathbb{E}_\theta[T(X)] = \tau(\theta), \quad \forall\ \theta.$$

   Calculs effectifs d'éstimateurs avec et sans biais par intégration de densité.
   – Estimateurs ponctuels classiques.
      – estimateur d'une moyenne :

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

      – Estimateur d'une proportion ou de la probabilité p d'un événement A :

$$\hat{p} = \frac{1}{n}(\text{nombre de fois que } X_i \text{ réalise } A).$$

      – Estimateur d'une variance

$$S_{n-1}^2 = \frac{1}{n-1}((X_1-\bar{X})^2 + \cdots + (X_n-\bar{X})^2).$$

      (On admettra que $\mathbb{E}[S_{n-1}^2] = \sigma^2$ et donc que S_{n-1}^2 est un estimateur sans biais de σ^2).

– Par convention, on utilise des lettres majuscules pour les v.a. et des lettres minuscules pour des observations particulières de ces variables. On utilise aussi la convention $\hat{\tau}$,$\hat{p}$,..., pour estimer des quantités τ, p...
– Définition générale de l'intervalle de confiance d'une quantité τ au risque α ou au seuil de confiance $1-\alpha$:

$$\mathbb{P}_\theta\big(T_{\min}(X) \leq \tau(\theta) \leq T_{\max}(X)\big) \geq 1-\alpha, \quad \forall\, \theta$$

où $T_{\min}(X)$ et $T_{\max}(X)$ sont des estimateurs.
– Intervalle de confiance de la moyenne μ lorsque l'écart-type est inconnu (le cas où l'écart-type σ_0 est connu ne sera pas traité)

$$\mathbb{P}\Big(\bar{X} - t_\alpha \frac{S_{n-1}}{\sqrt{n}} \leq \mu \leq \bar{X} + t_\alpha \frac{S_{n-1}}{\sqrt{n}}\Big) \geq 1-\alpha,$$

où t_α est l'écart d'une loi de Student $\mathcal{T}(n-1)$ à $n-1$ ddl au risque α, soit $t_\alpha = q_{1-\alpha/2}$ et $\mathbb{P}(|\mathcal{T}(n-1)| > t_\alpha) = \alpha$.
– Intervalle de confiance d'une proportion p

$$\mathbb{P}\Big(\hat{p} - z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_\alpha \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\Big) \geq 1-\alpha,$$

où z_α est tel que $\mathbb{P}(|\mathcal{N}(0,1)| > z_\alpha) = \alpha$. Utilisation des abaques.

11. **Intervalle de confiance II**
– Intervalle de confiance de la différence des moyennes : Cas de deux échantillons appariés

$$\overline{\Delta X} = \frac{1}{n}(\Delta X_1 + \cdots + \Delta X_n), \quad \Delta S_{n-1} = \Big(\frac{1}{n-1}\sum_{i=1}^n (\Delta X_i - \overline{\Delta X})^2\Big)^{1/2}.$$

$$\mathbb{P}\Big(\overline{\Delta X} - t_\alpha \frac{\Delta S_{n-1}}{\sqrt{n}} \leq \Delta\mu \leq \overline{\Delta X} + t_\alpha \frac{\Delta S_{n-1}}{\sqrt{n}}\Big) \geq 1-\alpha,$$

où t_α est tel que $\mathbb{P}(|\mathcal{T}(n-1)| > t_\alpha) = \alpha$.
– Intervalle de confiance de la différence des moyennes : cas de deux échantillons indépendants

$$\bar{X} = \frac{1}{n_A}(X_1 + \cdots + X_{n_A}), \quad \bar{Y} = \frac{1}{n_B}(Y_1 + \cdots + Y_{n_B}),$$

$$S_{AB}^2 = \frac{1}{n_A + n_B - 2}\Big(\sum_{i=1}^{n_A}(X_i - \bar{X})^2 + \sum_{i=1}^{n_B}(Y_i - \bar{Y})^2\Big).$$

$$\mathbb{P}\Big(\bar{X} - \bar{Y} - t_\alpha S_{AB}\sqrt{\tfrac{1}{n_A}+\tfrac{1}{n_B}} \leq \mu_A - \mu_B \leq \bar{X} - \bar{Y} + t_\alpha S_{AB}\sqrt{\tfrac{1}{n_A}+\tfrac{1}{n_B}}\Big) \geq 1-\alpha\,,$$

où t_α est tel que $\mathbb{P}(|\mathcal{T}(n_A + n_B - 2)| > t_\alpha) = \alpha$.

12. **Exercices de révision** Savoir absolument utiliser les fonctions statistiques de la calculette.

13. **Introduction aux tests d'hypothèses I**

- Hypothèse nulle (principale, préférentielle) H_0, hypothèse complémentaire (alternative, contre hypothèse) H_1, règle de décision, zone de rejet (ou zone critique) $\mathcal{R}$, zone d'acceptation $\mathcal{A}$), variable de décision, valeur critique.
- Risque de première espèce $\alpha = \mathbb{P}(\mathcal{R}|H_0)$ ou risque de rejeter H_0 à tort, de seconde espèce $\beta = \mathbb{P}(\mathcal{A}|H_1)$.

	$X \in \mathcal{A}$: on accepte H_0	$X \in \mathcal{R}$: on rejette H_0
en réalité : $\theta \in H_0$	prévision correcte	risque de première espèce
en réalité : $\theta \in H_1$	risque de seconde espèce	prévision correcte

- Un exemple parmi la leçon suivante comme support (par exemple, test d'une proportion ou test de la moyenne) .
- Retenir la méthodologie d'un test :
  (a) Donner le nom du test.
  (b) Définir l'hypothèse nulle H_0, bilatérale ou unilatérale. Un calcul numérique intermédiaire permet de définir un H_0 plus judicieux.
  (c) Ecrire la zone de rejet $\mathcal{R}$ correspondant à H_0. Rapeler les définitions des estimateurs entrant dans la définition de $\mathcal{R}$.
  (d) Choisir un seuil de confiance $1-\alpha$ ou niveau d'erreur α et calculer la valeur critique de la variable de décision au vue des données.
  (e) Conclure : accepter ou rejeter H_0 à l'erreur près α. Détailler la réponse.
  (f) Calculer la p-valeur du test : c'est-à-dire l'erreur que l'observateur commet en rejetant H_0 à tort au vue des données expérimentales.
- **Deuxième contrôle continu (30 mn)** Le contrôle portera sur les intervalles de confiances.

14. **Tests d'hypothèse usuels gaussiens II**
- Test d'une proportion p (cas des grandes valeurs de n) :

$$H_0 = \{p < p_0\}, \quad H_1 = \{p \geq p_0\}, \quad \mathcal{R} = \left\{\hat{p} - p_0 \geq q_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}\right\}$$
$$p_{val} = 1 - \mathbb{P}\Big(\mathcal{N}(0,1) < \frac{p_{obs} - p_0}{\sqrt{p_0(1-p_0)/n}}\Big)$$
$$H_0 = \{p = p_0\}, \quad H_1 = \{p \neq p_0\}, \quad \mathcal{R} = \left\{|\hat{p} - p_0| \geq z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}\right\}$$
$$p_{val} = 1 - 2 \times \mathbb{P}\Big(\mathcal{N}(0,1) < \frac{|p_{ops} - p_0|}{\sqrt{p_0(1-p_0)/n}}\Big)$$

où z_α et $q_{1-\alpha} = z_{2\alpha}$ sont tels que

$$\mathbb{P}(|\mathcal{N}(0,1)| > z_\alpha) = \mathbb{P}(\mathcal{N}(0,1) > q_{1-\alpha}) = \alpha.$$

– Test d'une moyenne μ d'écart-type inconnue :

$$H_0=\{\mu<\mu_0\},\quad H_1=\{\mu\geq\mu_0\},\quad \mathcal{R}=\Big\{\bar{X}\geq\mu_0+q_{1-\alpha}\frac{S_{n-1}}{\sqrt{n}}\Big\}$$
$$p_{val}=1-\mathbb{P}\Big(\mathcal{T}(n-1)<\frac{\mu_{obs}-\mu_0}{S_{n-1}/\sqrt{n}}\Big)$$
$$H_0=\{\mu=\mu_0\},\quad H_1=\{\mu\neq\mu_0\},\quad \mathcal{R}=\Big\{|\bar{X}-\mu_0|\geq t_\alpha\frac{S_{n-1}}{\sqrt{n}}\Big\}$$
$$p_{val}=\mathbb{P}\Big(|\mathcal{T}(n-1)|>\frac{|\mu_{obs}-\mu_0|}{S_{n-1}/\sqrt{n}}\Big)$$

où t_α et $q_{1-\alpha}=t_{2\alpha}$ sont tels que

$$\mathbb{P}(|\mathcal{T}(n-1)|\geq t_\alpha)=\mathbb{P}(\mathcal{T}(n-1)\geq q_{1-\alpha})=\alpha.$$

– Utiliser les tables statistiques pour calculer z_α et t_α. Remarquer que $z_{2\alpha}=q_{1-\alpha}$ et $t_{2\alpha}=q_{1-\alpha}$ pour les quantiles $q_{1-\alpha}$ respectifs des lois normales et de Student.

– Comparaison de deux moyennes (échantillons appariés) :

$$\overline{\Delta X}=\frac{1}{n}\sum_{i=1}^n\Delta X_i\quad\text{et}\quad \Delta S_{n-1}=\sqrt{\frac{1}{n-1}\sum\nolimits_{i=1}^n(\Delta X_i-\overline{\Delta X})^2}$$
$$H_0=\{\Delta\mu=0\},\quad H_1=\{\Delta\mu\neq0\},\quad \mathcal{R}=\Big\{|\overline{\Delta X}|\geq t_\alpha\frac{\Delta S_{n-1}}{\sqrt{n}}\Big\}$$
$$p_{val}=\mathbb{P}\Big(|\mathcal{T}(n-1)|>\frac{|\Delta\mu_{obs}|}{S_{n-1}/\sqrt{n}}\Big)$$

où $t_\alpha=q_{1-\alpha/2}$ est tel que $\mathbb{P}(|\mathcal{T}(n-1)|\geq t_\alpha)=\alpha$.

– Comparaison de deux moyennes (échantillons indépendants) :

$$S_{AB}=\sqrt{\frac{(n_A-1)S_A^2+(n_B-1)S_B^2}{(n_A-1)+(n_B-1)}},\quad H_0=\{\mu_A=\mu_B\}$$
$$H_1=\{\mu_A\neq\mu_B\},\quad \mathcal{R}=\Big\{|\bar{X}_A-\bar{X}_B|\geq t_\alpha S_{AB}\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\Big\}$$
$$p_{val}=\mathbb{P}\Big(|\mathcal{T}(n-1)|>\frac{|\bar{X}_{A,obs}-\bar{X}_{B,obs}|}{S_{AB}\sqrt{1/n_A+1/n_B}}\Big)$$

où t_α est tel que $\mathbb{P}(|\mathcal{T}(n_A+n_B-2)|\geq t_\alpha)=\alpha$. u $z_\alpha=q_{1-\alpha/2}$ est tel que $\mathbb{P}(|\mathcal{N}(0,1)|\geq z_\alpha)=\alpha$.

15. **Exercices de révision** Révision portant sur les tests d'hypothèse usuels paramétriques gaussiens (proportion, moyenne, comparaison).

16. **Test du chi-deux d'ajustement**

– Ajustement ou adéquation à une loi discrète à valeurs ou de modalités dans un ensemble fini, $\{\xi_1,\xi_2,\cdots,\xi_r\}$ et de probabilité $(p_1^0,\cdots,p_r^0)$,

$$\mathbb{P}(X=\xi_1)=p_1^0,\quad \mathbb{P}(X=\xi_2)=p_2^0,\quad\cdots\quad \mathbb{P}(X=\xi_r)=p_r^0.$$

– Efectifs théoriques np_j^0, effectifs observés n_j d'estimateur N_j égal au nombre de fois que la variable X_i prend la valeur ξ_j :

$$H_0 = \{p_j = p_j^0\}, \quad H_1 = \{p_j \neq p_j^0\}, \quad \mathcal{R} = \{D_{r-1}^2 \geq q_{1-\alpha}\},$$
$$\hat{D}_{r-1}^2 = \sum_{j=1}^{r} \frac{(N_j - np_j^0)^2}{np_j^0},$$

où D suit la loi du chi-deux à $r-1$ ddl et $q_{1-\alpha}$ est tel que $\mathbb{P}(\chi^2(r-1) \geq q_{1-\alpha}) = \alpha$.

– **Troisième contrôle continu (30 mn)** Contrôle portant sur les tests usuels gaussiens.

17. **Test du chi-deux d'indépendance**

– Table de contingence de deux v.a. X et Y de modalités $(\xi_1, \cdots, \xi_r)$ et $(\eta_1, \cdots, \eta_s)$, c'est-à-dire une table donnant les effectifs observés du couple $N_{i,j}$ et les effectifs marginaux correspondants,

	$\boldsymbol{\eta_1}$	$\boldsymbol{\eta_2}$	$\cdots$	$\boldsymbol{\eta_j}$	$\cdots$	$\boldsymbol{\eta_s}$	
$\boldsymbol{\xi_1}$	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1s}	$N_{1\cdot}$
$\boldsymbol{\xi_2}$	N_{21}	N_{22}	$\cdots$	N_{2j}	$\cdots$	N_{2s}	$N_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{\xi_i}$	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{is}	$N_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{\xi_r}$	N_{r1}	N_{r2}	$\cdots$	N_{rj}	$\cdots$	N_{rs}	$N_{r\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	$\cdots$	$N_{\cdot j}$	$\cdots$	$N_{\cdot s}$	n

– Effectifs observés $n_{j,k}$ du couple (X, Y) d'estimateur $N_{j,k}$ égal au nombre de fois que $X_i = \xi_j$ et $Y_i = \eta_k$. Distribution empirique de (X, Y), $\hat{p}_{j,k} = N_{j,k}/n$, distributions empiriques marginales de X et de Y : $\hat{p}_{j\cdot} = N_{j\cdot}/n$ et $\hat{p}_{\cdot k} = N_{\cdot k}/n$.

– Hypothèse nulle H_0 : dans tous les cas, c'est l'indépendance théorique $\hat{p}_{j,k} = \hat{p}_{j\cdot}\hat{p}_{\cdot k}$ ou $N_{j,k} = N_{j\cdot}N_{\cdot k}/n$.

– Table des effectifs théorique au cas où H_0 serait réalisée, c'est-à-dire une table donnant $N_{j\cdot}N_{\cdot k}/n$ dans chaque cas,

	$\eta_{\mathbf{1}}$	$\cdots$	$\eta_{\mathbf{j}}$	$\cdots$	$\eta_{\mathbf{s}}$
$\xi_{\mathbf{1}}$	$N_{1\cdot}N_{\cdot 1}/n$	$\cdots$	$N_{1\cdot}N_{\cdot j}/n$	$\cdots$	$N_{1\cdot}N_{\cdot s}/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\xi_{\mathbf{i}}$	$N_{i\cdot}N_{\cdot 1}/n$	$\cdots$	$N_{i\cdot}N_{\cdot j}/n$	$\cdots$	$N_{i\cdot}N_{\cdot s}/n$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\xi_{\mathbf{r}}$	$N_{r\cdot}N_{\cdot 1}/n$	$\cdots$	$N_{r\cdot}N_{\cdot j}/n$	$\cdots$	$N_{r\cdot}N_{\cdot s}/n$

– Test du chi-deux d'indépendance :

$$H_0 = \ll X \text{ et } Y \text{ sont indépendants } \gg, \quad \mathcal{R} = \{D^2_{(r-1)(s-1)} \geq q_{1-\alpha}\},$$
$$D^2_{(r-1)(s-1)} = \sum_{j=1}^{r}\sum_{k=1}^{s} \frac{(N_{j,k} - N_{j.}N_{.k}/n)^2}{N_{j.}N_{.k}/n}$$

$D^2_{(r-1)(s-1)}$ suit la loi du chi-deux à $(r-1)(s-1)$ ddl et le quantile $q_{1-\alpha}$ est tel que $\mathbb{P}(\chi^2((r-1)(s-1)) \geq q_{1-\alpha}) = \alpha$.

18. **Exercices de révision** Révision portant sur les tests d'hypothèse non paramétriques (indépendance et Chi-deux).
19. **TP machine I :** Introduction au maniement du logicien de statistique `R`. Les étudiants peuvent choisir de travailler en binôme ou en trinôpme. Les notes de TP machine font partie de la note de contrôle continu. Ce premier TP n'est pas à rendre.
20. **TP machine II :** Modélisation probabiliste. Le Tp montre dans une première partie comment créer un échantillon de loi discrète donnée à l'avance. Il montre dans une deuxième partie comment varie la vitesse de convergence dans le théorème central limite pour un échantillon de loi de Bernoulli.
21. **TP machine III :** Statistique inférentielle et intervalles de confiance. Le TP montre dans une première partie comment représenter par des histogrammes et des boxplots des fichiers de données. Il montre dans une deuxième partie comment calculer un intervalle de confiance, d'abords en utilisant les formules du cours, puis en utilisant les fonctions clef-en-main de `R`.
22. **TP machine IV :** Statistique inférentielle et tests d'hypothèse. Le TP montre comment analyser un fichier de données récupérées sur internet contenant le poids de 200 pots de caramel et de chocolat. Il montre d'abords comment réaliser un test de comparaison de moyenne entre les pots de caramel et les pots de chocolat, puis comment réaliser un test d'ajustement des poids des pots de chocolat à une loi normale $\mathcal{N}(\bar{x}, s^2_{n-1})$.