Consignes pour le choix des risques : dans tout l'examen
Pour la construction d'un intervalle de dispersion on choisira un risque $\alpha = 3.2$ %.
Pour la construction d'un intervalle de confiance on choisira un risque $\alpha = 3.8$ %.

Exercice 1 :

Le taux d'abstention aux élections européennes de 2004 a été en France de 57 %. Beaucoup d'analystes politiques pensent que l'abstention devrait augmenter lors des prochaines élections de juin 2009.

1. Par un sondage de 800 personnes, proposer une démarche statistique qui permette de corroborer ou pas ce pessimisme.

   Préciser la question posée, la loi de la variable aléatoire associée, les hypothèses, la loi de la statistique de test sous les hypothèses nulle et alternative, la zone et la borne de rejet ...

2. Imaginer un résultat de sondage appartenant à la zone de rejet (avec une erreur de 1re espèce de 0.005). Proposer alors une estimation ponctuelle et par intervalle du taux d'abstention à venir.

On observe 800 réalisations d'un VA de Ber(p) : "s'abstenir aux élections européennes" (OUI/NON) ;
$H_0 : p = .57$ "le taux d'abstention sera identique à celui de 2004
$H_1 : p > .57$ "le taux d'abstention sera supérieur à celui de 2004
La statistique de test est $\bar{X}$
de loi $\mathcal{N}(.57, .57 * .43/800)$ sous H_0
et de loi $\mathcal{N}(p, p * (1-p)/800)$ sous H_1
Le test statistique sera : "rejet de H_0 si $\bar{X} > .57 + l_\alpha * \sqrt(.57 * .43/800)$
Si $\alpha = 0.005$ alors $l_\alpha = 2.5758$; la Borne de rejet = 0.6150857.
Si le résultat du sondage est de 62% d'abstention prévue, on rejettera H_0
.62 sera une estimation ponctuelle de p
et l'intervalle de confiance de p à $\alpha = 3.8$ % ($l_{\alpha/2} = 2.0749$)
[$\bar{X} - l_{\alpha/2}\sqrt(\bar{X} * (1-\bar{X})/800); \bar{X} + l_{\alpha/2}\sqrt(\bar{X} * (1-\bar{X})/800)$]
soit
.62 - 2.0749 * sqrt(.62 * .38 / 800) = 0.5843926 et
.62 + 2.0749 * sqrt(.62 * .38 / 800) = 0.6556074

Dans la suite les tests statistiques seront construits avec une erreur de 1re espèce de 1 %.

Exercice 2 :

Chaque année, on réalise une enquête auprès d'étudiants de master sur les résultats de leur recherche de stage. La question est la suivante : "Sur les 5 premières demandes de stage (candidature spontanée par lettre) que vous avez envoyées, combien d'entre elles ont entraîné une réponse de l'entreprise (positive ou non) ? "

Jusqu'en 2008, il est admis que la distribution du "Nombre de réponses" est une loi binomiale de paramètres "n=5" et "p=0.25" et est donné par :

Nombre de réponses	**0**	**1**	**2**	**3**	**4**	**5**
Probabilité	0.2373	0.3955	0.2637	0.0879	0.0146	0.000977

En 2009, certains analystes pensent que cette hypothèse de loi n'est plus vérifiée à cause des effets de la crise. Pour cela, on interroge un échantillon de 500 étudiants de master sur les résultats de leur recherche de stage.

Proposer une démarche statistique complète qui permette de valider cet opinion.

Parmi les réponses des étudiants de master on relève 150 "0 réponse" et 240 "1 réponse". Utiliser cette information pour commencer vos calculs et peut-être proposer une conclusion.

On construit ici un test d'adéquation à la loi Bin(n=5, p=0.25) :
H_0 la variable "Nombre de réponses" suit une loi Bin(n=5, p=0.25)
H_1 la variable "Nombre de réponses" ne suit pas une loi Bin(n=5, p=0.25)
La statistique de test est alors $\sum_{r=0}^{5}(n_r - \tilde{n}_r)^2/\tilde{n}_r$
où par exemple $\tilde{n}_0 = 500 * 0.2373 = 118.65$ et $\tilde{n}_1 = 500 * 0.3955 = 197.75$
La Borne de rejet sera (associé à un ddl de 5 et $\alpha = 0.01$) égale à 15.0863
Pour les "0 réponse" et "1 réponse" les contributions sont :
$(150 - 118.65)^2/118.65 = 8.283375$
$(240 - 197.75)^2/197.75 = 9.026865$
La somme de ces 2 contributions égale à 17.31024 est déjà supérieure à la borne de rejet : on rejettera donc l'hypothèse H_0

Exercice 3 :

On élabore, à l'aide de plusieurs questions, un score (compris entre 0 et 100) sur la capacité d'un étudiant à se positionner sur le marché de l'emploi : plus le score est proche de 0, plus l'étudiant aura une capacité réduite à trouver de l'emploi ; à l'inverse plus le score est proche de 100, plus l'étudiant aura des facilités à trouver de l'emploi.

Depuis le début des années 2000, il était établi que les paramètres d'espérance et de variance de cette variable score étaient de 60 et 81 respectivement.

1. Lorsqu'en 2004 on interrogeait 121 étudiants, dans quel intervalle se trouvait alors très probablement le score moyen de ces 121 étudiants.
2. Pensant que la crise financière de 2009 a rendu plus difficile et plus longue la recherche d'emploi à la fin des études universitaires, proposer une démarche qui mette en évidence cette idée. Mener les calculs en supposant que vous interrogerez 225 étudiants en cette fin d'année universitaire.

   La question qui suit est indépendante de ce qui précède

3. Ce score est maintenant élaboré sur 160 étudiants et on a noté aussi un facteur géographique à 3 modalités.

   Notant s_{ij} le score du j^e étudiant dans la localisation i ; les résultats sont regroupés dans le tableau suivant :

Localisation	Nbre d'étudiants	Moyenne par localisation
1	$n_1 = 30$	$\bar{s}_{1.} = 51.7$
2	$n_2 = 50$	$\bar{s}_{2.} = 59.413$
3	$n_3 = 80$	$\bar{s}_{3.} = 61.105$

   et on donne aussi :

$$\sum_{i=1}^{3}\sum_{j=1}^{n_i}(s_{ij} - \bar{s}_{i.})^2 = 13438$$

   Construire une démarche statistique permettant de mettre en évidence que le score de l'étudiant change avec la localisation.

   Identifier clairement le modèle, les hypothèses mises en jeu, les paramètres inconnus, ainsi que les estimations de ces paramètres sous chacune des hypothèses.

1. C'est un intervalle de dispersion puisque pour 2004 on connaît la loi de $\bar{X}$ (moyenne des scores sur 121 étudiants) : $\mathcal{N}(60, \sigma^2 = 81/121)$:
   $[\ 60 - l_{\alpha/2}\sqrt(81/121); 60 + l_{\alpha/2}\sqrt(81/121)\ ]$
   avec $l_{\alpha/2} = 2.1444$ et $\sqrt(81/121) = 9/11$
   soit [58.24549 ; 61.75451]
2. On construit ici un test statistique de l'hypothèse :
   H_0 " $\mu = 60$ " le score construit est identique (en espérance)
   contre
   H_1 " $\mu < 60$ " le score construit diminue (en espérance)
   Le test statistique sera : "rejet de H_0 si $\bar{X} < 60 - l_\alpha * \sqrt(81/225)$
   avec $l_\alpha = 2.3263$; $\sqrt(81/225) = 9/15 = 3/5 = 0.6$
   la borne de rejet sera alors : 58.60422

3. La démarche statistique permettant de mettre en évidence que le score de l'étudiant change avec la localisation est l'Analyse de la Variance (ANOVA).

   On observe 30 + 50 + 80 = 160 variables S_{ij} indépendantes , d'espérance μ_i (i=1,2 ou 3) et on construit le test de

   H_0 " $\mu_1 = \mu_2 = \mu_3$ " les scores (en espérance) sont identiques pour chaque localisation

   contre

   H_1 " non H_0 " les scores (en espérance) ne sont pas identiques pour chaque localisation

   La statistique de test est $V_{Inter}/V_{Intra} * (n-3)$ (n=160) et la borne de rejet est (ddl = 2) : 9.2103

   On prendra dans ce qui suit V pour variation ou dispersion : on peut faire la même chose pour V comme variance en divisant par n

   Pour le calcul de V_{Inter} on a besoin de la moyenne générale :

   $\bar{s}_{..} = (n_1 * \bar{s}_{1.} + n_2 * \bar{s}_{2.} + n_3 * \bar{s}_{3.})/n$

   soit :

   $\bar{s}_{..} = (30 * 51.7 + 50 * 59.413 + 80 * 61.105)/160 = 58.81281$

   $V_{Inter} = n_1 * (\bar{s}_{1.} - \bar{s}_{..})^2 + n_2 * (\bar{s}_{2.} - \bar{s}_{..})^2 + n_3 * (\bar{s}_{3.} - \bar{s}_{..})^2$

   $30 * (51.7 - 58.81281)^2 + 50 * (59.413 - 58.81281)^2 + 80 * (61.105 - 58.81281)^2 = 1956.104$

   V_{Intra} est donnée par l'énoncé = 13438

   Donc la valeur de la statistique de test est : 1956.104 / 13438 * 157 = 22.85372

   On rejette donc H_0 (les scores (en espérance) sont identiques pour chaque localisation)

   Puisque on rejette H_0 les paramètres inconnus μ_1, μ_2, μ_3 seront estimés respectivement par $\bar{s}_{1.}, \bar{s}_{2.}, \bar{s}_{3.}$.

   Si on ne rejetait pas H_0 , le paramètre inconnu μ (un seul) est estimé par la moyenne générale.

   Réponses supplémentaires posible

   L'hypothèse forte de l'Anova est l'homogénéité des variances (un seul paramètre σ^2 qui est estimé sous H_1 par la variance Intra : 13438/160 = 83.9875

   Sous H_0, le paramètre σ^2 serait estimé par la variance totale.