BTS - Statistique Inférentielle - Cours

1. Résultats préliminaires

1.1. Introduction

But : On étudie un caractère réel (par exemple : taille, énergie, ...) sur une population en cherchant, à partir d'un échantillon de cette population, à retrouver les indicateurs du caractère (moyenne, écart-type, ...).

Inférence : Opération logique par laquelle on admet une proposition en vertu de sa liaison avec d'autres propositions déjà tenues pour vraies (le Robert).

1.2. Rappels

Propriété 1 : Rappels :

- Lorsque X est une variable aléatoire :
  - $\mathtt{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ et $\sigma(X) = \sqrt{\mathtt{Var}(X)}$ (variance et écart-type)
- si a et b sont deux constantes :
  - $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$
  - $\mathtt{Var}(aX + b) = a^2\mathtt{Var}(X)$
- Lorsque $S = X_1 + \cdots + X_n$ est une somme de variables aléatoires :
  - $\mathbb{E}(S) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n)$: les espérances s'ajoutent ;
  - $\mathtt{Var}(S) = \mathtt{Var}(X_1) + \cdots + \mathtt{Var}(X_n)$: les variances s'ajoutent **si les variables aléatoires sont indépendantes**.

Exercice 1 : Montrer que si X est une variable aléatoire d'espérance μ et d'écart-type σ, alors $Z = \dfrac{X - \mu}{\sigma}$ est une variable aléatoire d'espérance 0 (on dit centée) et d'écart-type 1 (on dit réduite).
Exprimer X en fonction de Z et compléter (a, b sont des constantes): $a \leqslant Z \leqslant b \Leftrightarrow \ldots\ldots\ldots \leqslant X \leqslant \ldots\ldots\ldots$

Exercice 2 : $X_1, \ldots, X_n$ sont des variables aléatoires indépendantes. On note $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$.

1. Montrer que $\mathbb{E}\left(\overline{X}\right) = \frac{1}{n}(\mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n))$ et que $\mathtt{Var}\left(\overline{X}\right) = \frac{1}{n^2}(\mathtt{Var}(X_1) + \cdots + \mathtt{Var}(X_n))$
2. En déduire que si tous les X_i ont même moyenne μ, même variance v et donc même écart-type σ, on a :
   - $\mathbb{E}\left(\overline{X}\right) = \mu$;
   - $\mathtt{Var}\left(\overline{X}\right) = \dfrac{v}{n}$; en déduire que $\sigma\left(\overline{X}\right) = \dfrac{\sigma}{\sqrt{n}}$
3. Montrer que si les X_i suivent une loi de Bernoulli ($X = 1$ avec une probabilité p, sinon $X = 0$), alors :
   - $\mathbb{E}(S) = np$ et $\sigma(S) = \sqrt{np(1-p)}$: en effet, une variable aléatoire suivant une loi binomiale est une somme de variables aléatoire indépendantes suivant une loi de Bernoulli de même paramètre p.
   - $\mathbb{E}\left(\overline{X}\right) = p$
   - $\sigma\left(\overline{X}\right) = \sqrt{\dfrac{p(1-p)}{n}}$

2. Modèle

Définition 1 :

- *Une **population** de taille N est modélisée par un ensemble de variables aléatoires $\{Y_1, \ldots, Y_N\}$, qui représentent le caractère mesuré sur chaque individu.*
  On suppose, sauf cas particulier, que ces variables aléatoires suivent la même loi et sont indépendantes.

- *On sélectionne n individus, qui forment un **échantillon** de taille n<N ; on a donc un ensemble de variables aléatoires $\{X_1, \ldots, X_n\}$, tel que, par exemple, $X_1 = Y_3$, $X_2 = Y_{14}$, ...*

- *Une **réalisation** de cette échantillon consiste à donner à chaque variable aléatoire de l'échantillon une valeur réelle, selon la loi suivie par cette variable aléatoire : on mesure le caractère étudié sur l'échantillon.*
  Une réalisation d'un n-échantillon se traduit donc par l'obtention de n valeurs : $x_1, \ldots, x_n$ (notées en minuscule).

- *On appelle **estimateur** sur un échantillon de taille n fonction (à valeurs réelles) de n variables $h(x_1, \ldots, x_n)$; par exemple $h(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \cdots + x_n)$ (formule de la moyenne, notée $\overline{x}$ en statistiques), ou bien $h(x_1, \ldots, x_n) = x_1 \times x_2$ (peu d'utilité).*

- *$h(X_1, \ldots, X_n)$ étant une variable aléatoire, on peut noter $\mathbb{E}(f)$, si elle existe, l'**espérance** de l'estimateur f.*

- *Par essence, l'estimateur vise à approcher un paramètre de la population (par exemple la moyenne du caractère observé sur la population, une proportion, ...).*
  *Ainsi, lorsque k est la notation du paramètre approché, on **peut noter** l'estimateur $\widehat{k}$ (avec un accent circonflexe), lorsqu'il n'y a pas d'ambigüité.*

- *Pour mesurer l'erreur entre l'estimation et la réalité, on utilise le **biais** $\mathbb{B}$, défini par $\mathbb{B}\left(\widehat{k}\right) = \mathbb{E}\left(\widehat{k}\right) - k$.*
  *Lorsque son biais est nul, on dit que l'estimateur est **sans biais**.*

Propriété 2 : L'estimateur $\widehat{\overline{x}} = \frac{1}{n}(x_1 + \cdots + x_n)$ est sans biais.
Démonstration : On note μ la moyenne sur la population entière.

$$\mathbb{B}\left(\widehat{\overline{X}}\right) = \mathbb{E}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) - \mu = \frac{1}{n}\left(\mathbb{E}\left((X_1\right) + \cdots + \mathbb{E}\left(X_n\right)\right) - \mu = \frac{1}{n}(\mu + \cdots + \mu) - \mu = \frac{1}{\cancel{n}}\cancel{n}\mu$$

Exercice 3 : On note $\overline{X}$ la variable aléatoire définie par $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ et on note V la variable aléatoire définie par $V = \overline{X^2} - \overline{X}^2$.
On suppose que $X_1, \ldots, X_n$ ont toutes pour espérance μ et pour variance v.
Démontrer que $\mathbb{E}(V) = \dfrac{n-1}{n}v$; (utiliser $\mathbb{E}(T^2) = \mathrm{Var}(T) + \mathbb{E}(T)^2$) ;
en déduire que le calcul de la variance sur l'échantillon ne fournit pas un estimateur sans biais de la variance sur la population ;
en déduire qu'un estimateur sans biais de la variance sur la population est donné par $\dfrac{n}{n-1}V$.

Définition 2 : Écart-type ponctuel (ou corrigé)
*On utilise l'**écart-type ponctuel ou corrigé** $s_n = \sqrt{\frac{n}{n-1}}\sigma_n$ comme estimateur de l'écart-type sur la population globale ; σ_n étant l'écart-type calculé sur un échantillon de taille n.*

3. Intervalle de confiance

On ne connaît pas la valeur moyenne µ d'un caractère observé sur la population.
Pour calculer l'**intervalle de confiance** I_c de la moyenne µ **de risque** α (ou **de confiance 1-**α) à partir d'un échantillon E de taille n :

confiance 1-α	0,99	0,98	0,95	0,90
risque α	0,01	0,02	0,05	0,10
z	2,58	2,33	1,96	1,65

Méthode 1 : Intervalle de confiance d'une moyenne

1. on détermine la valeur z telle que $P(-z \leqslant Z \leqslant z) = 1 - \alpha$, ($Z$ désignant une variable aléatoire suivant une loi normale centrée réduite) ;
2. on calcule la moyenne $\overline{x}$ sur l'échantillon E.
   - si l'on connaît l'écart-type σ sur la population : $\boxed{I_c = \left[\overline{x} - z\frac{\sigma}{\sqrt{n}} ; \overline{x} + z\frac{\sigma}{\sqrt{n}}\right]}$
   - si l'on ne connait pas l'écart-type sur la population, on utilise l'écart-type ponctuel s_n (corrigé) et : $\boxed{I_c = \left[\overline{x} - z\frac{s_n}{\sqrt{n}} ; \overline{x} + z\frac{s_n}{\sqrt{n}}\right]}$

Exercice 4 : Intervalle de confiance d'une moyenne
Une machine produit des tubes dont la longueur doit être fixée. On prélève 100 tubes dans la production : On s'intéresse à la longueur moyenne µ des tubes sur l'ensemble de la production.

Longueur	[994;998[	[998;1002[	[1002;1006[
Nombre	26	70	4

1. Déterminer un intervalle de confiance au risque de 5%.
2. Déterminer un intervalle de confiance au risque de 1%.
3. 1000mm est-il une moyenne réaliste ?

On ne connaît pas la valeur p d'une proportion observée sur la population.
Pour calculer l'**intervalle de confiance** I_c de la proportion p, **de risque** α (ou **de confiance 1-**α) à partir d'un échantillon E de taille n :

Méthode 2 : Intervalle de confiance d'une proportion

1. on détermine la fréquence f observée sur l'échantillon E ;
2. lorsque p n'est pas très proche de 0 ou 1 et que n⩾30, on utilise la loi normale centrée réduite pour déterminer la valeur z telle que $P(-z \leqslant Z \leqslant z) = 1 - \alpha$, sinon on utilise une loi binomiale (cf exercice) ;
3. $\sqrt{\frac{f(1-f)}{n-1}}$ étant une estimation ponctuelle de l'écart-type sur la population on a : $\boxed{I_c = \left[f - z\sqrt{\frac{f(1-f)}{n-1}} ; f + z\sqrt{\frac{f(1-f)}{n-1}}\right]}$

Exercice 5 : Intervalle de confiance d'une proportion
D'après un sondage sur n=2501 personnes, $\widehat{p} = 51\%$ souhaitent voter pour le candidat A. On note p la proportion de personnes votantes pour A dans la population.

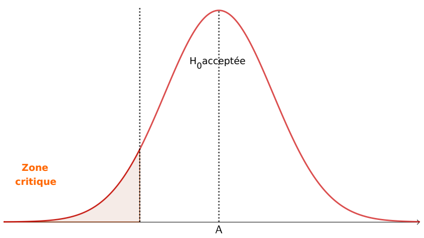
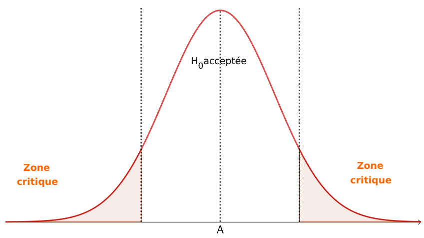
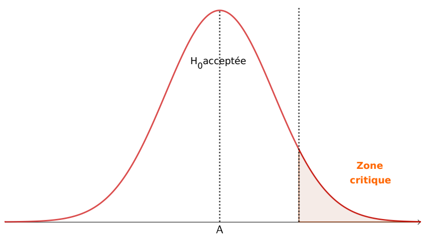
1. Déterminer un intervalle de confiance au risque de 5%.
2. Le candidat A est-il presque sûr (à 95%) d'être élu ?
3. Déterminer la taille n qu'aurait dû avoir l'échantillon si la réponse à la question précédente est négative

4. Tests de validité d'hypothèse

On veut tester l'hypothèse qu'une certaine valeur a, existante mais non connue, sur une population donnée, correspond bien à une valeur fixée A, ou bien a «changé».

Méthode 3 :

- On prélève un échantillon de taille n dans la population.
- On énonce l'hypothèse nulle, notée H_0, qui correspond à une situation «inchangée» : $H_0 : a = A$.
- On détermine l'hypothèse alternative, notée H_1, qui est l'hypothèse que l'on peut montrer avec le test : 3 possibilités.
- On fixe un niveau de confiance/risque et on détermine l'intervalle de test, noté I_t, qui correspond à l'intervalle de confiance de risque α sauf dans le cas d'un test unilatéral où on remplace une de ses bornes par l'infini.
  L'extérieur de cet intervalle est appelé **zone critique**.
- On applique la **règle de décision** qui suit :
  - si $\widehat{a}$, obtenu sur l'échantillon, appartient à I_t, **on accepte H_0 au niveau de confiance** $1-\alpha$;
  - sinon $\widehat{a}$ est dans la zone critique, et **on rejette H_0 au seuil** α.

test unilatéral à gauche	**test bilatéral**	**test unilatéral à droite**
$H_1 : a \leqslant A$	$H_1 : a \neq A$	$H_1 : a \geqslant A$
$I_t = [A-h ; +\infty[$	$I_t = [A-h ; +A+h]$	$I_t = [-\infty ; A+h[$
		

Exercice 6 : Proposer un test visant à vérifier si le rythme cardiaque ralentit suite à un don du sang.

Définition 3 : Erreur de première et deuxième espèce ; puissance *(lors d'un test, il y a un risque de se tromper)*

- *l'erreur de **première espèce** α correspond au risque de rejeter H_0 alors qu'elle est vraie (faux positif).*
- *l'erreur de **deuxième espèce** β correspond au risque d'accepter H_0 alors qu'elle est fausse (faux négatif).*

Décision \ réalité	***H_0 vraie***	***H_1 vraie***
H_0 acceptée	*bonne décision* $(1-\alpha)$	*risque β 2nde espèce*
H_1 acceptée	*risque α 1re espèce*	*bonne décision* $(1-\beta)$ *=puissance*

- *La **puissance du test** est le risque de rejeter H_0 alors qu'on doit en effet rejeter H_0.*

Exercice 7 : Test bilatéral relatif à une proportion : Jeu de pile ou face
Pour vérifier qu'une pièce est bien équilibrée (non truquée), on jette n=100 fois cette pièce et on note X la variable aléatoire comptant le nombre de pile obtenus ; le but de l'exercice est de construire un test.

1. Formuler l'hypothèse nulle H_0 correspondant à une pièce bien équilibrée ; quelle est l'hypothèse alternative H_1 dans le cas d'un test bilatéral.
2. Calculer un intervalle de confiance au risque α de 5%, en récisant la zone critique, et en utilisant :
   - une loi binomiale
   - une loi normale ; y a-t-il réellement une différence ?
3. Énoncer la règle de decision.
4. On sait qu'il existe sur le marché des pièces truquées qui donnent pile dans 2 cas sur 3. Déterminer le risque β de seconde espèce (accepter que la pièce ne soit pas truquée sachant qu'elle l'est) et donner la puissance du test.
5. Recommencer dans le cas d'un dé (X compte le nombre de 6), on fixe le test à 50 lancers. Y a-t-il une différence entre l'utilisation de la loi normale et binomiale pour l'intervalle de fluctuation, dans ce cas ?