ENSTA ParisTech 1ère année
Statistiques (MA101), 2017-2018
christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

# TD6: Intervalles de confiance, tests

**Exercice 1.**
Le chiffre d'affaires moyen d'un commerçant, calculé sur les trente derniers jours, est de 2000 euros, avec un écart type de valeur $s_{30} = 300$ euros.

1. Si on admet que son chiffre d'affaires quotidien peut être représenté par une v.a. $X$ de loi normale de moyenne $m$ et de variance $\sigma^2$ inconnues, donner un intervalle de confiance de niveau 0.95 pour le paramètre $m$.

2. Le commerçant affirme que son chiffre d'affaire quotidien peut être représenté par une gaussienne d'espérantce 1850 €. Que répondez-vous avec ces données? Quelle critique éventuelle pouvez-vous formuler sur la méthodologie?

3. Obtient-on le même intervalle si $\sigma$ est connu et de valeur $\sigma = 300$ ?

***Correction.*** *Posons $n = 30$, notons $X_i$ le chiffre d'affaires du jour $i$, notons $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, et $\widehat{\sigma}^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$.*

1. *Si les v.a. $X_i$ sont i.i.d. $\mathcal{N}(m\ ;\ \sigma^2)$ alors en utilisant le fait que la loi de $T := \frac{\sqrt{n}(\overline{X}-m)}{\sqrt{\widehat{\sigma}^2}}$ est la loi de Student $\mathcal{T}(n-1)$, et donc la relation $\mathbb{P}(-2,045 \leq T \leq 2,045) = 95\%$, on obtient :*

$$\mathbb{P}(\overline{X} - \frac{2,045\sqrt{\widehat{\sigma}^2}}{\sqrt{n}} \leq m \leq \overline{X} + \frac{2,045\sqrt{\widehat{\sigma}^2}}{\sqrt{n}}) = 95\%.$$

   *Le calcul numérique donne $I_{obs} = [1888\ ;\ 2112]$.*

2. *On utilise l'équivalence entre l'intervalle de Student et le test de Student. Ici, 1850 n'appartient pas à l'intervalle de confiance observé, donc on rejette le test de Student de $(H_0) : m = 1850$ contre $(H_1) : m \neq 1850$, avec un risque de première espèce $\alpha = 5\%$.*

   *Il faudrait d'une part s'assurer que les observations d'un jour à l'autre soient bien indépendantes. Par ailleurs, il est possible que le mois observé soit particulièrement bénéfique. Si le commerçant parle de la loi de son chiffre d'affaire quotidien en référence à l'année, il faudrait prendre l'échantillon au hasard sur toute l'année.*

3. *Dans ce cas on utilise la statistique $U := \frac{\sqrt{n}(\overline{X}-m)}{\sqrt{\sigma^2}}$ dont la loi est la loi $\mathcal{N}(0\ ;\ 1)$. La même technique que précédemment nous conduit à*

$$\mathbb{P}(\overline{X} - \frac{1,96\sqrt{\sigma^2}}{\sqrt{n}} \leq m \leq \overline{X} + \frac{1,96\sqrt{\sigma^2}}{\sqrt{n}}) = 95\%.$$

*Le calcul numérique donne* $I_{obs} = [1892,65\ ;\ 2107,35]$ *et on remarque que cet intervalle est plus précis que le précédent, puisqu'il n'y a plus d'incertitude sur la variance.*

**Exercice 2.**
Soient deux échantillons gaussiens $X = (X_1, \dots, X_n)$ et $Y = (Y_1, \dots, Y_m)$ indépendants et chacun iid de lois respectives $\mathcal{N}(\mu_X, \sigma_X^2)$ et $\mathcal{N}(\mu_Y, \sigma_Y^2)$. Les paramètres d'espérance et de variance sont inconnus.

1. Construire un intervalle de confiance bilatère de $\beta = \sigma_X^2/\sigma_Y^2$
2. Construire un test d'égalité des variances au niveau 5%. La pvalue observée est 0.01. Que conclure, et avec quel risque?

***Correction.*** *Sous* $(H_0)$*, les estimateurs non biaisés de la variance dans chacun des deux échantillons sont indépendants et*

$$Z = \frac{\widehat{\sigma}_X^2/\sigma_X^2}{\widehat{\sigma}_Y^2/\sigma_Y^2} \sim \mathcal{F}(n-1, m-1)$$

*On note* $qf_\alpha^{n-1,m-1}$ *le quantile d'ordre* $\alpha$ *d'une loi de Fisher de paramètres* $n-1$ *et* $m-1$*. On en déduit*

$$IC(\sigma_Y^2/\sigma_X^2) = \left[ qf_{\alpha/2}^{n-1,m-1} \frac{\widehat{\sigma}_Y^2}{\widehat{\sigma}_X^2};\ qf_{1-\alpha/2}^{n-1,m-1} \frac{\widehat{\sigma}_Y^2}{\widehat{\sigma}_X^2} \right]$$

*ou*

$$IC(\sigma_X^2/\sigma_Y^2) = \left[ qf_{\alpha/2}^{m-1,n-1} \frac{\widehat{\sigma}_X^2}{\widehat{\sigma}_Y^2};\ qf_{1-\alpha/2}^{m-1,n-1} \frac{\widehat{\sigma}_X^2}{\widehat{\sigma}_Y^2} \right]$$

*puisque* $qf_{1-\alpha}^{n-1,m-1} = 1/qf_\alpha^{m-1,n-1}$*. En effet, soit* $K \sim \mathcal{F}(n,m)$

$$\alpha = \mathbb{P}\left(K \leq qf_\alpha^{n,m}\right) = \mathbb{P}\left(\frac{1}{qf_\alpha^{n,m}} \leq \frac{1}{K}\right) = 1 - \mathbb{P}\left(\frac{1}{K} \leq \frac{1}{qf_\alpha^{n,m}}\right)$$

*soit* $1-\alpha = \mathbb{P}\left(\frac{1}{K} \leq \frac{1}{qf_\alpha^{n,m}}\right)$.

*Rem: attention, il y a des coquilles sur l'expression de l'IC dans le poly p. Les bonnes informations sont celle du corrigé de TD.*

*Le test bilatère de fonction de test*

$$\varphi_\alpha(X,Y) = 1 - \mathbb{1}_{qf_{\alpha/2} < Z < qf_{1-\alpha/2}}$$

*est de niveau* $\alpha$*.*

*La p-value étant inférieure au niveau, on rejette l'égalité avec un risque de* 5%*.*

ENSTA ParisTech 1ère année
Statistiques (MA101), 2017-2018
christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

**Exercice 3.**
Le jeu de données `birthwt`[1] enregistre les poids de naissance (en gr.) de nouveaux-nés pour 115 mères non fumeuses et 74 mères fumeuses.

1. La p-valeur du test de Shapiro-Wilks vaut 0.33 pour les mères non fumeuses et 0.42 pour les mères fumeuses. Quelle modélisation proposez-vous ?

2. Effectuer un test d'égalité des variances des deux échantillons au niveau 5%. On observe $\widehat{\sigma}^2_{nf} = 566492$ et $\widehat{\sigma}^2_f = 435118$. Le quantile à 0.975 de la loi $\mathcal{F}(114, 73)$ vaut 1.54, et celui de la loi $\mathcal{F}(73, 114)$ vaut 1/0.66. Quelle décision prenez-vous?

   Donner l'expression de la p-value. Est-elle supérieure ou inférieure à 5%

3. On observe $\bar{x}_{nf} = 3055.7$ et $\bar{x}_f = 2772$. Tester l'égalité des poids moyens des nouveaux-nés en fonction du statut fumeur/non fumeur de leur mère.

   Calculer la p-value du test. Donner un intervalle de confiance de la différence des poids moyens. Interpréter.

4. On suspecte que le tabac ait une influence négative sur le poids des nouveaux-nés, ceux de mère fumeuse étant plus chétifs. Que répondez-vous?

***Correction.*** *1. Le test de Shapiro-Wilks teste l'adéquation de l'échantillon à une loi gaussienne (hypothèse nulle) contre la non-adéquation (alternative). Pour les deux échantillons, la p-value est supérieure au niveau demandé pour le test, on conserve donc l'adéquation de chaque échantillon à la modélisation gaussienne, avec un risque de seconde espèce inconnu.*

*2. Sous* $(H_0) : \sigma_{nf} = \sigma_f$

$$Z = \frac{\widehat{\sigma}^2_{nf}}{\widehat{\sigma}^2_f} \sim \mathcal{F}(n_{nf} - 1, n_f - 1)$$

*où l'on utilise les estimateurs non biaisés de la variance dans chacun des échantillon. L'alternative est* $(H_1) : \sigma_{nf} \neq \sigma_f$. *La région de rejet*

$$\{Z > qf^{n_{nf}-1,n_f-1}_{1-\alpha/2}\} \cup \{Z < qf^{n_{nf}-1,n_f-1}_{\alpha/2}\}$$

*est de risque* $\alpha$

*A.N.:* $\widehat{\sigma}^2_{nf}/\widehat{\sigma}^2_f = 1.3 < qf^{115-1,74-1}_{1-\alpha/2} = 1.54$. *On conserve l'égalité des variances, avec un risque de seconde espèce inconnu.*

*La pvalue* $= 2(1 - \mathbb{P}_{\mathcal{F}(114,73)}(Z > \widehat{\sigma}^2_{nf}/\widehat{\sigma}^2_f))$ *est supérieure à 5% puis qu'on a conservé* $(H_0)$. *Le calcul avec un logiciel donne* 0.22.

[1] accessible par exemple à partir du package MASS du logiciel R

3. *Il s'agit du test de Student de comparaison de moyennes de deux échantillons gaussien indépendants. Puisqu'il s'agit de tester l'égalité, les hypothèses sont* $(H_0) : \mu_{nf} = \mu_f$ *contre* $(H_1) : \mu_{nf} \neq \mu_f$. *La statistique de test est*

$$T = \frac{\overline{X}_{nf} - \overline{X}_f}{\widehat{\sigma}\sqrt{\frac{1}{n_{nf}} + \frac{1}{n_f}}} \sim \mathcal{T}(n_f + n_{nf} - 2), \text{ avec } \widehat{\sigma}^2 = \frac{(n_f - 1)\widehat{\sigma}_f^2 + (n_{nf} - 1)\widehat{\sigma}_{nf}^2}{n_f + n_{nf} - 2}$$

*suit une loi de Student de paramètre* $n_f + n_{nf} - 2$. *Soit* $q_t^{1-\alpha/2}$ *le quantile d'ordre* $1-\alpha/2$ *d'une loi* $\mathcal{T}(n_f + n_{nf} - 2)$, *la région de rejet* $\{|T| > q_t^{1-\alpha/2}\}$ *de niveau exact* $\alpha$

*A.N.:* $t_{obs} = 2.65 > 1.97 = q_t^{1-\alpha/2}$, *on rejette donc l'égalité des poids avec un risque de première espèce* $\alpha = 5\%$

*L'intervalle de confiance bilatère de* $\mu_{nf} - \mu_f$

$$IC(\mu_{nf} - \mu_f) = \left[\overline{X}_{nf} - \overline{X}_f \pm q_t\, \widehat{\sigma}\sqrt{\frac{1}{n_{nf}} + \frac{1}{n_f}}\right]$$

*est de niveau exact* $1 - \alpha$.

*A.N.: Sa valeur observée sur les données vaut* $I_{obs} = [73; 495]$. 0 *n'y appartient pas, on rejette donc bien* $(H_0) : \mu_{nf} - \mu_{nf} = 0$.

4. *Il s'agit maintenant de tester* $(H_0) : \mu_f - \mu_{nf} = 0$ *contre* $(H_1) : \mu_{nf} - \mu_f > 0$. *Même statistique de test, même loi sous* $(H_0)$ *que dans la question précédente. La région de rejet est maintenant* $\{T > q_t^{1-\alpha}\}$ *de niveau exact* $\alpha$

*AN:* $t_{obs} = 2.65 > q_t^{1-\alpha/2} > q_t^{1-\alpha}$. *On rejette donc* $(H_0)$ *et on conclut à une baisse de poids pour les nouveaux nés dont les mères sont fumeuses. Attention, ce n'est qu'une liaison qui est montrée, pas une relation de cause à effet.*

**Exercice 4.**
A partir du jeu de données précédents, un indicateur de faible poids a été défini: il vaut 1 si le poids du nouveau-né est inférieur à un certain seuil, 0 sinon. 29 nourrissons sont de poids faible parmi les 115 non-fumeuses et 30 sont de poids faibles parmi les 74 fumeuses.

Tester l'hypothèse d'une liaison négative du tabac sur le poids des nouveaux-nés: décrire la(es) méthodologie(s) et conclure en précisant le risque de la décision prise.

***Correction.*** *C'est un test de comparaison de moyennes, mais dans le cadre d'échantillons de Bernoulli. Soient* $\pi_1$ *la probabilité pour un nourrisson d'être de faible poids*

*parmi les mères fumeuses, $\pi_0$ celle pour un nourrisson d'être de faible poids parmi les mères non fumeuses.*

*On teste $\pi_0 = \pi_1$ contre $\pi_1 > \pi_0$.*

*Sous $(H_0)$, on a la même espérance de la loi de Bernoulli, $\pi = \pi_1 = \pi_2$ donc la même variance $\sigma^2 = \pi(1-\pi)$ où $\pi$ est la probabilité (commune) pour un nourrisson d'être de faible poids. Or $\pi$ est inconnue, on l'estime par la proportion de nouveaux-nés de faible poids dans l'ensemble des deux échantillons $\widehat{\pi} = (n_1\widehat{\pi}_1 + n_0\widehat{\pi}_0)(n_1+n_0)$. Utilisons maintenant la statistique*

$$T_{n_0,n_1} = \frac{\widehat{\pi}_1 - \widehat{\pi}_0}{\widehat{\sigma}\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}} = \frac{\widehat{\pi}_1 - \widehat{\pi}_0}{\sqrt{\frac{n_0\widehat{\pi}_0+n_1\widehat{\pi}_1}{n_0+n_1}\frac{n_0(1-\widehat{\pi}_0)+n_1(1-\widehat{\pi}_1)}{n_0 n_1}}}$$

*En utilisant le TLC+ Slutsky, $T_{n_0,n_1}$ converge en loi vers une gaussienne centrée réduite. On en déduit la région de rejet de la forme*

$$\mathcal{R} = \{T_{n_0,n_1} > q_{0.95}^{\mathcal{N}(0,1)}\}$$

*qui est asymptotiquement de niveau $\alpha = 5\%$. Comme les proportions observées vérifient $n_j\widehat{\pi}_j(1-\widehat{\pi}_j) > 12$ pour $j = 1, 0$ on peut utiliser la loi asymptotique à distance finie, avec un niveau qui n'est qu'approximativement $\alpha$.*

*AN: $\widehat{\pi}_1 = 30/74 = 0.405$, $\widehat{\pi}_0 = 29/115 = 0.252$, $\widehat{\pi} = 59/189 = 0.31$, $t_{obs} = 2.21$, supérieur au quantile d'ordre 95% d'une loi gaussienne centrée réduite qui vaut 1.64, on rejette la non liaison, et on conclut avec un risque de 5% que la proportion de bébés de faible poids est plus important chez les mères fumeuses, d'où la liaison positive entre la prise de tabac et le faible poids de naissance.*

**Exercice 5.**
Déterminer un intervalle de confiance bilatère exact de l'espérance d'une loi exponentielle à partir d'un $n$-échantillon iid de cette loi.

Construire un IC asymptotique.

***Correction.*** *Soit $\mu = \mathbb{E}(X_i)$. On utlise la méthode du pivot. Nous avons vu dans les TDs précédent que la région de rejet $T = \sum_i X_i/\mu$ suit un loi Gamma $\Gamma(n,1)$. D'où*

$$\mathbb{P}(\gamma_n^{\alpha/2} < T < \gamma_n^{1-\alpha/2}) = 1-\alpha$$

*où $\gamma_n^\alpha$ est le quantile de $\Gamma(n,1)$ d'ordre $\alpha$. Soit*

$$IC(\mu) = \left[\sum_i X_i/\gamma_n^{1-\alpha/2}; \sum_i X_i/\gamma_n^{\alpha/2}\right]$$

*Par le TLC, on a, avec $q^*$ le quantile d'ordre $1-\alpha/2$ de la loi gaussienne centrée réduite*

$$\mathbb{P}\left[-q^* < \sqrt{n}\left(\frac{\bar{X}}{\mu} - 1\right) < q^*\right] \to 1-\alpha$$

*d'où l'IC de niveau asymptotique $\alpha$*

$$IC(\mu) = \left[\frac{\bar{X}}{1+q^*/\sqrt{n}};\ \frac{\bar{X}}{1-q^*/\sqrt{n}}\right]$$