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

TD5: Tests (suite)

Exercice 1.
Reprendre l'exercice 3 du TD 4.

1. Montrer que le test proposé dans cet exercice est UPP.
2. Donner l'expression de la p-value. Celle-ci vaut 0.5. Quelle est la décision?
3. Calculer la p-value du test asymptotique. Commenter

Correction. *1. La loi exponentielle d'espérance μ est de densité $f(x;\mu) = \frac{1}{\mu}\exp -(x/\mu)$. On pose l'hypothèse nulle $(H_0) : \mu = d_0$ contre l'alternative $(H_1) : \mu = \mu_1 < d_0$. La région de rejet optimale du test d'un test d'hypothèses simples de niveau α est celle de Neyman-Pearson*

$$\mathcal{R}_\alpha = \left\{ x \in \mathbb{R}^n; \frac{L(\mu_1;x)}{L(d_0;x)} > k_\alpha \right\}, \text{ avec } \mathbb{P}_{d_0}((X_1,\dots,X_n) \in \mathcal{R}_\alpha)) = \alpha$$

où $L(\mu;x)$ est la vraisemblance des observations, soit, comme les observations sont iid:

$$\mathcal{R}_\alpha = \left\{ x \in \mathbb{R}^n; \frac{d_0}{\mu_1} \exp\left((\frac{1}{d_0} - \frac{1}{\mu_1}) \sum_i x_i \right) > k_\alpha \right\}$$

La région critique s'exprime uniquement avec la statistique $S = \sum_i X_i$ et est de la forme $\{s \in \mathbb{R}; (\mu_1 - d_0)s > k'_\alpha\}$. Comme $\mu_1 < d_0$ la région de rejet UPP est tq $\mathbb{P}_{(H_0)}(\sum_i X_i/\mu_0 < k'_\alpha) = \alpha$. Or, $\sum_i X_i/d_0 \sim \Gamma(1,n)$ d'où $k'_\alpha = \gamma_{n,\alpha}$, où $\gamma_{n,\alpha}$ est le quantile d'ordre α d'une loi $\Gamma(1,n)$: $\mathbb{P}_{(H_0)}(\sum_i X_i/d_0 < \gamma_{n,\alpha}) = \alpha$.

2. Par définition, la pvalue $= \mathbb{P}_{(H_0)}(\sum_i X_i/d_0 < tobs)$ où t_{obs} est le résultat de la statistique $\sum_i X_i/d_0$ de loi $\Gamma(1,n)$ sous (H_0): $t_{obs} = \sum_i x_i/d_0 = 10 \times 1550/1600 = 9.69$. On a $\mathbb{F}_{\Gamma(1,n)}(t_{obs}) = 0.5 < 0.05 = \alpha$. On ne peut rejeter (H_0) au niveau 5%. On la conserve donc avec un rique de seconde espèce inconnu.

3. On peut récrire la p-value:

$$\begin{aligned}\mathbb{P}_{(H_0)}(\sum_i X_i/d_0 < t_{obs}) &= \mathbb{P}_{(H_0)}\left(\sqrt{n}\frac{\bar{X} - d_0}{d_0} < \sqrt{n}\,(t_{obs} - 1)\right) \\ &\simeq \mathbb{F}^*(\sqrt{n}(t_{obs} - 1))\end{aligned}$$

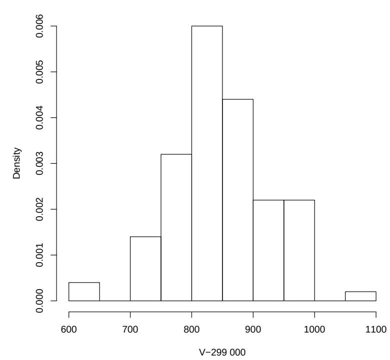
où $\mathbb{F}^*$ est la fonction de répartition de $\mathcal{N}(0,1)$ en approximant à distance finie la loi de $\bar{X}/d - 0 - d_0$ par celle de $\mathcal{N}(0,1)$ (TLC). A.N.: $\mathbb{F}^*(\sqrt{n}(t_{obs} - 1)) = \mathbb{F}^*(-0.099) = 0.46$. L'approximation est raisonnable et ne change pas la décision du test sur ce cas particulier.

christine.keribin@math.u-psud.fr, Ennio Fedrizzi, Jade Giguelay, Cyril Marzouk, Loïc Richier, Morgan Schmitz

Exercice 2.
On dispose de 100 mesures de la vitesse de la lumière (km s^{-1}) effectuées en 1879 par Michelson, lors de la célèbre expérience qui lui valut le prix Nobel de physique en 1907[1]. Avec les instruments de cette époque, les scientifiques avaient calculé une vitesse de la lumière avec une légère imprécision: $c_{1879} = 299840$ km/s. Aujourd'hui, la vitesse de la lumière est établie à $c_{2018} = 299792$ km/s.

Dans tout l'exercice, les niveaux des tests sont pris à 5%

1. Commenter l'histogramme des mesures de Michelson.



   Quelle modélisation poposer? La p-value du test de Shapiro-Wilks vaut 0.51. Quelle décision prendre, quel est son rique?

2. On souhaite tester l'hypothèse nulle que les mesures de Michelson sont conformes à la valeur de la vitesse de la lumière communément admise en 1879.

   Construire un test approprié.

   Sur l'échantillon de Michelson, la moyenne observée vaut $m = 299852$ et l'écart-type corrigé vaut $s = 79$. Quelle décision prendre et à quel risque?

   Calculer la p-value et vérifier cette décision.

3. Les observations de Michelson sont-elles en accord avec la valeur de la vitesse de la lumière mesurée aujourd'hui? Justifier la réponse par un test.

4. On souhaite tester la précision σ de la mesure. Définir la statistique de test, la région de rejet, la p-value, la décision et son risque dans les deux cas suivants:

   (a) $(H_0) : \sigma = 90$ contre $(H_1) : \sigma < 90$.

[1] John Rice: Mathematical Statistics and Data Analysis, Duxbury Press 1995

(b) L'écart-type σ de la mesure peut-il être considéré comme égal à 91 ? Commenter.

Correction. *1. L'histogramme pourrait être compatible avec celui de données iid gaussiennes. La pvalue du test de Shapiro-Wilks ((H_0) la loi est gaussienne) calculée par un logiciel permet de prendre une décision sans connaître directement les tables de probabilité de la statistique de test correspondante. Ici, pvalue $> 0.05 = \alpha$. On conserve (H_0) (échantillon gaussien) avec un risque de seconde espèce inconnue. On propose donc que les mesures sont issues d'un n-échantillon iid gaussien d'espérance μ et de variance σ^2 inconnues.*

2. On teste $(H_0) : \mu = c_{1879}$ contre $(H_1) : \mu \neq c_{1879}$. On utilise le test de Student dont la statistique $T = \sqrt{n}(\bar{X} - c_{1879})/\widehat{\sigma}$ (avec $\widehat{\sigma}^2$ l'estimateur non biaisé de la variance) suit sous (H_0) une loi de Student $\mathcal{T}(n-1)$. La région de rejet du test bilatéral est $\mathbb{P}_{(H_0)}(|T| > q_{1-\alpha/2})$ où $q_{1-\alpha/2}$ est le quantile de la loi $\mathcal{T}(n-1)$ d'ordre $1-\alpha/2$.

A.N.: $t_{obs} = \sqrt{n}(m - c_{1879})/s = 1.57$ et $|t_{obs}| < 1.98 = q_{0.975}$. On conserve la valeur c_{1879}. On prend un risque (de seconde espèce) inconnu en prenant cette décision.

On a pval $= \mathbb{P}_{(H_0)}(|T| > t_{obs}) = 2\mathbb{P}_{(H_0)}(T < -|t_{obs}|) = 2\mathbb{F}_{\mathcal{T}(n-1)}(-1.57)) = 0.12 > \alpha = 5\%$, même décision.

3. Même test de Student avec $(H_0) : \mu = c_{2018}$ contre $(H_0) : \mu \neq c_{2018}$, la statistique de test est maintenant $T = \sqrt{n}(\bar{X} - c_{2018})/\widehat{\sigma}$

A.N.: $t_{obs} = 7.64 > q_{0.975}$, pval $= 1.374e-11 < 5\% = \alpha$. On rejette ($H_0$) au risque de première espèce α. La p-value étant très faible, le test est hautement significatif pour dire que la valeur de 1879 n'est pas en accord avec celle d'aujourd'hui.

4. Test de la variance d'un échantillon gaussien iid. La statistique de test (basée sur $\widehat{\sigma}^2$ l'estimateur non biais de la variance) est $K = (n-1)\widehat{\sigma}^2/\sigma_0^2$ et suit sous (H_0) une loi du Khi-deux à (n-1) ddl.

(a) Région de rejet unilatérale à gauche de niveau α:

$$\mathbb{P}[K < q^{(\alpha)}_{\chi^2(n-1)}] = \alpha$$

A.N.: $\sigma_0 = 90$; $k_{obs}/90^2 = (100-1) \times (79/90)^2 = 76.3$, pval $= \mathbb{P}(K < k_{obs}) = \mathbb{F}_{\chi^2(100-1)}(k_{obs}) = 0.04$. On rejette ($H_0$). La variance est inférieure à valeur 90 avec un risque de première espèce $\alpha = 5\%$.

(b) *Région de rejet bilatérale de niveau α:*

$$\mathbb{P}[\{K < q^{(\alpha/2)}_{\chi^2(n-1)}\} \cup \{q^{(1-\alpha/2)}_{\chi^2(n-1)} < K\}] = \alpha$$

A.N.: $\sigma_0 = 91$; $k_{obs}/91^2 = (100-1) \times (79/91)^2 = 74.6$, $pval = 2\mathbb{P}(K < k_{obs}) = 2\mathbb{F}_{\chi^2(100-1)}(k_{obs}) = 0.90$. On conserve la valeur 91 avec un risque de seconde espèce inconnu.

Ceci n'est pas incohérent avec le résultat précédent, les hypothèses n'étant pas définies sur les mêmes intervalles d'une part, et les deux décisions ne sont pas prises avec les mêmes risques d'autre part.

Exercice 3.
On souhaite tester si une contamination ayant eu lieu dans une région y a augmenté le taux de prématurés, qui est normalement de 5%. Dans la région contaminée, il y a eu sur un an 1200 naissances dont 80 prématurées.

1. Construire un test UPP de niveau $\alpha = 1\%$.
2. En déduire un test asymptotique. Indiquer la décision et son risque.
3. Donner l'expression de la p-value. Sans faire de calcul, pouvez-vous dire si elle est supérieure ou inférieure à α?
4. Calculer la fonction puissance.

Correction. *1. Etapes de construction:*

- *Modèle: Le résultat observé sur un individu est binaire: on considère donc un échantillon iid $Z_1, \ldots, Z_n \sim \mathcal{B}(1, \pi)$, avec $\pi = \mathbb{E}(Z_1)$*
- *Hypothèses: $(H_0) : \pi = \pi_0 = 5\%$ contre $(H_1) : \pi > \pi_0$*
- *le théorème de Neyman-Pearson pour deux hypothèses simples donne une région de rejet de la forme*

$$\left\{ \frac{\pi_1^{\sum_i Z_i}(1-\pi_1)^{n-\sum_i Z_i}}{\pi_0^{\sum Z_i}(1-\pi_0)^{n-\sum_i Z_i}} > k_\alpha \right\} = \left\{ \left(\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} \right)^{\sum_i Z_i} > k_\alpha \right\}$$

soit, en prenant en compte le fait que $\pi_1 > \pi_0$, on cherche $\tilde{k}_\alpha$ tel que $\mathcal{R} = \{\sum_i Z_i > \tilde{k}_\alpha\}$ soit de niveau α. $\sum_i Z_u$ suit sous (H_0) une loi binomiale $\mathcal{B}(n, \pi_0)$, d'où k_α est le quantile d'ordre $1-\alpha$ de cette binomiale.

2. *Comme $n\pi_0 > 10$, $n(1-\pi_0) > 10$, on peut approcher la binomiale par une loi normale $\mathcal{N}(n\pi_0, n\pi_0(1-\pi_0))$, d'où la statistique de test*

$$T_n = \frac{\bar{Z} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \underset{(H_0)}{\overset{appr}{\sim}} \mathcal{N}(0,1)$$

Région de rejet : $\mathcal{R} = \{T_n > q^*_{1-\alpha}\}$ où $q^*_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de $\mathcal{N}(0,1)$, ($q^*_{0.99} = 2.32$). $\mathbb{P}_{H_0}(\mathcal{R}) \simeq \alpha$.

Décision: $t_{obs} = (80/1200 - 0.05)/\sqrt{0.05 \times 0.95/1200} = 2.64$, $t_{obs} > qnorm(0.99) = 2.33$. *On rejette la non augmentation et on conclut avec un risque (de première espèce) de 1% à l'augmentation du taux de naissances prématurées.*

3. *Par définition, $pval = \mathbb{P}_{(H_0)}(T_n > t_{obs})$. Le test ayant rejeté (H_0), on sait que $pval < \alpha = 1\%$.*

4. *Puissance en fonction de π_1*

$$\begin{aligned}
\mathbb{P}_{\pi_1}\left(\frac{\bar{Z} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} > q^*_{1-\alpha}\right) &= \mathbb{P}_{\pi_1}\left(\frac{\bar{Z} - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}} > \sqrt{\frac{\pi_0(1-\pi_0)}{\pi_1(1-\pi_1)}} q^*_{1-\alpha} + \frac{\pi_0 - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}}\right) \\
&\simeq 1 - F^*\left(\sqrt{\frac{\pi_0(1-\pi_0)}{\pi_1(1-\pi_1)}} q^*_{1-\alpha} + \frac{\pi_0 - \pi_1}{\sqrt{\frac{\pi_1(1-\pi_1)}{n}}}\right)
\end{aligned}$$

où F^* est la fonction de répartition d'une loi gaussienne centrée réduite.

Exercice 4.

Soit $X_1, \ldots, X_n$ un n-échantillon de loi uniforme sur l'intervalle $[0,\ \theta]$, $\theta > 0$.

Soit $Y = \max(X_1, \ldots, X_n)$.

On considère le test de

$$H_0 :\ \theta = \frac{1}{2} \ \text{ contre } \ H_1 :\ \theta > \frac{1}{2}\ ,$$

et on décide de rejeter H_0 si $Y > c$.

1. Montrer que la fonction de distribution de Y est $F(y) = \theta^{-n} y^n$ si $y \in [0, \theta]$, 0 si $y \leq 0$, 1 si $y \geq \theta$.

2. Définir la fonction de puissance $\pi_c(\theta)$ du test proposé. Expliciter son expression.

3. Déterminer c pour que le test proposé soit de niveau α.

   Quelle est sa puissance lorsque $n \to \infty$?

4. Quelle doit être la taille de l'échantillon n pour obtenir une puissance de 0.98 sous l'alternative $\theta = 3/4$?

5. On observe un échantillon de taille $n = 20$ et une valeur de Y de 0.48 : calculer la p-valeur du test.

   Quelle est la décision du test? Quel est le risque d'erreur associé à cette décision?

Correction. *1. La densité de $\mathcal{U}(0,1)$ est $f(x) = \mathbb{1}_{x\in[0;\theta]}/\theta$ et sa fonction de répartition $F(x) = x/\theta$ pour $x \in [0;\theta]$, nulle pour $x \leq 0$, et valant 1 pour $x \geq \theta$. En utilisant l'indépendance des observations X_i on a, pour tout x:*

$$\mathbb{P}(Y \leq x) = \mathbb{P}\left(\cap_{i=1}^n (X_i \leq x)\right) = \prod_{i=1}^n \mathbb{P}(X_i \leq x) = F^n(x).$$

Remarque: $Y = \widehat{\theta}$ est l'estimateur du maximum de vraisemblance de θ, qui est asymptotiquement sans biais et consistant: en effet, la densité h de Y est, par dérivation en x

$$h(x) = nF^{n-1}(x)f(x) = \frac{n}{\theta^n}x^{n-1}\mathbb{1}_{[0;\theta]}(x)$$

d'où

$$\mathbb{E}(\widehat{\theta}) = n\theta^{-n}\int_0^\theta x^n dx = \frac{n}{n+1}\theta; \quad \mathrm{Var}(\widehat{\theta}) = \frac{n}{(n+2)(n+1)^2}\theta^2$$

et

$$R(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) + [Biais(\widehat{\theta})]^2 = \frac{2\theta^2}{(n+1)(n+2)}$$

2. La puissance π_c est définie sur $]1/2, +\infty[$ par

$$\forall \theta \in]1/2, +\infty[\qquad \pi_c(\theta) = \mathbb{P}_\theta(Y > c) = \begin{cases} 1 - c^n\theta^{-n} & si \quad c \in [0,\theta], \\ 1 & si \quad c < 0, \\ 0 & si \quad c > \theta. \end{cases}$$

3. On choisit c tel que $\alpha = \mathbb{P}_{1/2}(Y > c) = 1-(2c)^n$, c'est-à-dire $c = (1-\alpha)^{1/n}/2$. Pour tout $\theta \in]1/2, +\infty[$, on a alors

$$\pi_c(\theta) = 1 - (1-\alpha)/(2\theta)^n \underset{n\to+\infty}{\longrightarrow} 1.$$

4. *On cherche* n *tel que* $\pi_c(3/4) \geqslant 0.98$*, c'est à dire* $0.95/(3/2)^n \leqslant 0.02$*. D'où*

$$n \geqslant \frac{\ln(0.95) - \ln(0.02)}{\ln(3) - \ln(2)} \approx 9.597.$$

   Il faut donc prendre $n = 10$*.*

5. *La p-valeur est* $pval = P_{H_0}(Y \geqslant 0.495) = 1 - (2 \times 0.495)^{20} \approx 0.182 > 0.05$*. On ne rejette pas* H_0 *au niveau* 5%*. On ne connaît pas le risque (de seconde espèce) de se tromper en prenant cette décision.*