Université de Poitiers
Département de Mathématiques

2M13–Statistique inférentielle
Master MMAS, Année 2008–09

ESTIMATION, INTERVALLES DE CONFIANCE DE LA MOYENNE D'UNE LOI NORMALE

Rappels

On considère un espace probabilisé $(E, \mathcal{E}, P)$ où la mesure de probabilité P est mal connue. Cet espace est vu comme l'espace d'état d'une variable aléatoire X muni de la loi de cette variable. Pour déterminer certaines caractéristiques ou paramètres de cette loi, on effectue une suite de mesures ou de réalisations de la variable, ce qui conduit à un échantillon : suite d'individus $(\omega_1, \ldots, \omega_n)$, suite de variables aléatoires $(X_1, \ldots, X_n)$ de même loi P (souvent supposées indépendantes), suite d'observations $(x_1, \ldots, x_n)$ de la variable X.

On cherche à estimer un paramètre de cette loi qui est souvent un réel λ. Pour cela, on définit un estimateur $\Lambda = f(X_1, \ldots, X_n)$ (variable aléatoire) qui doit posséder de bonnes propriétés (consistance, absence de biais, entre autres). Il y correspond une estimation $\ell = f(x_1, \ldots, x_n) \in \mathbf{R}$. Une telle estimation est souvent appelée « estimation ponctuelle ».

Nous avons vu grâce à des simulations que la convergence vers la moyenne par exemple (loi forte des grands nombres) peut être lente, voire même impossible lorsque la moyenne à estimer n'est pas définie (lois de Cauchy, ...). Ainsi, si on ne dispose que d'un nombre restreint d'observations, l'estimation peut être assez erronée. Il faut donc se donner des marges d'erreur.

Soient $(X_1, \ldots, X_n)$ un échantillon et $\alpha \in]0, 1[$. Un intervalle de confiance de seuil de risque α ou de niveau de confiance $1 - \alpha$ du paramètre $\lambda \in \mathbf{R}$ de la loi P est la donnée de deux variables $\Lambda^- = f^-(X_1, \ldots, X_n) \leq \Lambda^+ = f^+(X_1, \ldots, X_n)$ telles que $\mathbf{P}\{\lambda \in [\Lambda^-, \Lambda^+]\} = 1 - \alpha$. Ceci ne dit pas grand chose. Généralement, on souhaite que l'estimateur se trouve dans l'intervalle et que celui-ci soit cohérent avec le problème posé (le paramètre appartient à $\mathbf{R}$, $\mathbf{R}_+$, $[0, 1]$). Bien souvent, pour des problèmes de lourdeur de calcul, on ne détermine que des intervalles de confiance approximatifs, autrement dit, tels que $\mathbf{P}\{\lambda \in [\Lambda^-, \Lambda^+]\} \approx 1 - \alpha$. Pour cela on utilise des théorèmes de limite en loi (souvent le théorème central limite).

1. Intervalles de confiance de la moyenne d'une loi normale

Nous considérons une variable X de loi $\mathcal{N}(\mu, \sigma^2)$, donc de loi normale de moyenne μ et de variance σ^2 ($E = \mathbf{R}$ et $\mathcal{E} = \mathcal{B}(\mathbf{R})$). Nous cherchons à estimer μ à l'aide d'un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes toutes de loi $\mathcal{N}(\mu, \sigma^2)$. Le premier cas est celui où σ est connu (ce qui est assez rare à mon avis). L'intervalle de confiance qu'on choisit alors est

$$\left[\overline{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\,,\, \overline{X}_n + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi Normale $\mathcal{N}(0, 1)$. Lorsque σ n'est pas connu, on considère

$$\left[\overline{X}_n - t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\,,\, \overline{X}_n + t_{1-\alpha/2}\frac{S_n}{\sqrt{n}}\right]$$

où $t_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi de Student à $n - 1$ degrés de liberté.

Exercice 1 (non pratique). — Vérifier que ce sont dans les deux cas des intervalles de confiance exacts.

Remarque. — On rappelle que $\overline{X}_n = (X_1 + \cdots + X_n)/n$ est la moyenne observée et $S_n^2 = ((X_1^2 + \cdots + X_n^2) - \overline{X}_n^2)/(n-1)$ est la variance échantillon.

Supposons que les données `x` contenues dans une table `A` correspondent à un échantillon d'un loi normale dont les deux paramètres sont inconnus. Si on se fixe un seuil $\alpha = 0.05$ par exemple, le logiciel `SAS` calcule aisément l'intervalle de confiance de la moyenne cherché :

```
proc means data=A alpha=0.05 clm mean std;
     var x;
run;
```

L'option `clm` (pour *confidence limits*) demande le calcul de l'intervalle de confiance. Les options `mean` et `std` demande l'affichage de la moyenne et de la variance échantillon observées. Si les données se présentent sous la forme données/fréquence, il faut le préciser à l'aide de l'instruction `freq` dans le corps de la procédure. Comme toujours, plus de précisions peuvent être obtenue avec « l'aide ».

Exercice 2. — On a tiré un échantillon de 375 personnes parmi des hommes de 50 à 69 ans, mesuré leur taux de cholestérol X et obtenu après regroupement des données (les valeurs données pour x correspondent aux centres des intervalles de regroupement) :

x	90	110	130	150	170	190	210	230	250	270	290	310	330	350	370	390	410	430	450
n	2	6	11	50	79	60	58	35	39	20	7	3	2	2	0	0	0	0	1

(i) Calculer la moyenne observée, l'écart-type échantillon et tracer un histogramme de ces données au moyen de la procédure `univariate` (voir « l'aide »). La distribution du taux de cholestérol observé vous paraît-elle correspondre à une loi normale ?

(ii) Précisez pourquoi il semble légitime de déterminer un intervalle de confiance pour la moyenne par l'une des formules précédentes. Calculer un tel intervalle lorsque $\alpha = 0.05$ et lorsque $\alpha = 0.01$.

(iii) Comparer les résultats obtenus avec `SAS` avec ceux que vous pouvez obtenir avec votre calculatrice en utilisant l'une des expressions proposées.

2. Simulations élémentaires d'une loi $\mathcal{N}(\mu, \sigma^2)$

La signification des intervalles de confiance n'est pas très immédiate. Il se peut qu'avec un échantillon on obtienne une estimation très proche de la réalité (*a priori* cachée), ou bien au contraire qu'on s'en soit beaucoup écarté. Afin de matérialiser cette notion, nous allons procéder à des simulations.

La commande `rannor(0)` génère pseudo-aléatoirement des nombres selon la loi Normale $\mathcal{N}(0,1)$. Pour le voir, l'exercice suivant propose de générer 1000 valeurs obtenues avec celle-ci et d'en tracer l'histogramme.

Exercice 3. — Saisir et exécuter le programme suivant :

```
title "Étude d'une série de 1000 sorties de rannor(0)";

data simulation1;
     n=1000;
     do i=1 to n;
          u=rannor(0);
```

```
          output;
     end;
     drop n i;
run;

proc print data=simulation1 (obs=50);
run;

/* quelques param\‘etres graphiques pour l’histogramme... */

pattern v=l1 c=blue;
symbol color=blue width=2;

proc univariate data=simulation1;
     histogram u/normal(mu=0 sigma=1 color=red w=2);
run;
quit;
```

Après avoir exécuté ce programme, regarder attentivement les sorties : graphiques et surtout texte. Consulter la documentation pour en savoir plus sur les différentes commandes ou procédures employées.

EXERCICE 4. — Nous désirons simuler maintenant une variable de loi $\mathcal{N}(\mu, \sigma^2)$ avec $\mu = 5$ et $\sigma = 2$. Reprendre le programme précédent pour ce faire. Comparer les sorties des deux programmes.

3. **Simulations d'intervalles de confiance**

Nous considérons la loi normale de moyenne 5 et d'écart-type 2 et prendrons pour seuil $\alpha = 0.05$. En supposant $\sigma = 2$ connu, nous tirons un échantillon de taille 25 par exemple. Celui-ci nous permet d'avoir une estimation par intervalle de μ (qu'on sait valoir 5). Cet intervalle peut contenir ou non le paramètre à estimer : tout dépend de l'échantillon obtenu. Par conséquent, nous allons considérer 100 échantillons et examiner la proportion des cas où le paramètre μ est dans l'intervalle de confiance.

EXERCICE 5. — (i) Formuler par écrit la définition des intervalles de confiance que nous considérons.

(ii) Adapter le programme précédent pour ce problème, et commenter :

```
data simulation3;
     n=25; nechantillon=100; m=5; sigma=2;
     do j=1 to nechantillon;
          do i=1 to n;
               x=sigma*rannor(0)+mu;
               output;
          end;
     end;
     drop i nechantillon;
run;
```

(iii) Nous allons calculer les moyennes échantillon (noter que par défaut les données utilisées sont les dernières sorties). Commenter.

```
proc means noprint;
     var x;
     output out=moyennes mean=m;
     by j;
run;
```

(iv) Nous allons reprendre la dernière table ou ensemble de données créée pour construire l'ensemble des intervalles de confiance au seuil $\alpha = 0.05$ pour lequel on a $z_{1-\alpha/2} = z_{0.975} \approx 1.9600$.

```
data intervalles;
     set moyennes; /* acquisition des donn\'ees de la table "moyennes"*/
     sigma=2; n=25;
     a=m-sigma*1.96/sqrt(n);
     b=m-sigma*1.96/sqrt(n);
run;

title "Bornes des intervalles de confiance";
proc print data=intervalles;
     var j m a b;
run;
```

Retrouver le premier intervalle par un calcul direct. Quel est le nombre d'intervalles ne contenant pas le paramètre μ ?

(v) Nous allons essayer de représenter graphiquement les intervalles de confiance. La solution proposée n'est peut-être pas la plus élégante.

```
data bornesinf;
     set moyennes;
     sigma=2; n= 25;
     borne=m-sigma*1.96/sqrt(n);
run;

data bornessup;
     set moyennes;
     sigma=2; n= 25;
     borne=m+sigma*1.96/sqrt(n);
run;

data intervalles;
     set bornesinf bornessup;
     mu=5;
run;

/* Quelques param\‘etres graphiques...*/

symbol1 i=hilo c=red;
symbol2 i=join c=blue width=2;

title "Intervalles de confiance de mu=5 au niveau alpha=5%"
prog glot data=intervalles;
     plot borne*j mu*j/overlay;
run;
quit;
```

Débuguer si nécessaire... et retrouver les résultats précédents.

Exercice 6. — Profiter du temps restant pour créer un fichier TEX (ou LATEX) où insérer les graphiques obtenus au format `eps` ainsi que les sorties texte pertinentes.

Université de Poitiers
Département de Mathématiques

2M13–Statistique inférentielle
Master MMAS, Année 2008–09

ESTIMATION D'UNE PROPORTION, TESTS CLINIQUES

1. **Estimation d'une proportion**

Dans une certaine population, on cherche à estimer la proportion des individus ayant une certaine propriété. Pour cela, on choisit un échantillon de manière judicieuse, puis on calcule la proportion observée. Bien sûr, cette estimation ponctuelle a une pertinence relative. La personne menant l'étude ne peut se prononcer que sur ce qu'il/elle a vu et proposer un résultat sous la forme d'un intervalle de confiance.

Lorsque l'échantillonnage est réalisé à la façon d'un tirage avec remise de taille n, on s'attend légitimement à voir le nombre d'individus parmi possédant la propriété étudiée se répartir selon une loi binomiale $\mathcal{B}(n,\pi)$ où $\pi \in [0,1]$ est la proportion cherchée. Les calculs avec des lois binomiales sont rapidement très lourds.

Exercice 1. — Compter le nombre de jeunes filles en séance de TP parmi les étudiants. À l'aide de l'abaque ou de la table en fin de document, déterminer un intervalle de confiance de la proportion correspondante au seuil $\alpha = 0{,}05$. Commenter.

Même si les capacités de calcul des ordinateurs courants ne cessent de croître, la détermination d'intervalles de confiance exacts pour une proportion n'est qu'assez rarement pertinente lorsque la population étudiée est suffisamment vaste et qu'on peut disposer de grands échantillons. Une proportion π à évaluer est alors souvent estimée par intervalles approchés du type

$$\left[p - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\,,\, p + z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right]$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale $\mathcal{N}(0,1)$, p la proportion observée et n la taille de l'échantillon.

Exercice 2. — (i) Justifier la pertinence d'une telle définition pour l'intervalle de confiance approché.

(ii) Comparer les résultats donnés par l'abaque ou la table avec celui donné par cette formule asymptotique dans le cadre de l'exercice précédent (on pourra se servir de sas en créant un tableau à deux valeurs et en utilisant la fonction `probit` pour la détermination du quantile [voir aussi `probnorm` ainsi que d'autres fonctions du même type]).

2. **Intervalles de confiance de la différence de deux proportions dans une étude prospective**

Pour étudier l'incidence d'un facteur d'exposition sur l'apparition d'une maladie (ou au contraire sa guérison), on considère deux populations : E la population des sujets exposés et E^c ou $\overline{E}$ la population des sujets non exposés. Dans ces deux populations les probabilités respectives de contraction de la maladie envisagée sont notées π_1 et π_2. Pour estimer ces probabilités, on considère deux échantillons de ces deux populations et, après un certain temps, on évalue les proportions respectives de malades de chancun de ces échantillons. Ce type d'étude est appelée *prospective*, ou *étude de cohortes* ; c'est ce que l'on rencontre dans les

essais cliniques. Les probabilités π_1 et π_2 sont qualifiées de risques. On désire les comparer : le risque atttribuable est $\varrho\alpha = \pi_1 - \pi_2$ et le risque relatif est $\varrho\varrho = \pi_1/\pi_2$.

Les résultats d'une telle étude peuvent être présentés sous la forme d'un tableau

	Malades (M)	Non malades (M^c)	Total
Exposés (E)	a	b	$n_1 = a + b$
Non exposés (E^c)	c	d	$n_2 = c + d$
Total	$a + c$	$b + d$	$n = n_1 + n_2$

L'estimation de π_1 est donnée par a/n_1, celle de π_2 par c/n_2. De même, $\varrho\alpha$ est éstimé par $ra = p_1 - p_2$ et $\varrho\varrho$ par $rr = p_1/p_2$. On admet que lorsque a, b, c, d sont tous suffisamment grands (supérieurs à 10 en général), les intervalles

$$\left[p_1 - p_2 - z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\,,\; p_1 - p_2 + z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right]$$

et

$$\left[\ln(rr) - z_{1-\alpha/2}\sqrt{\frac{1-p_1}{a} + \frac{1-p_2}{c}}\,,\; \ln(rr) + z_{1-\alpha/2}\sqrt{\frac{1-p_1}{a} + \frac{1-p_2}{c}}\right]$$

sont des intervalles de confiance asymptotiques (ou approximatif) de $\varrho\alpha$ et $\ln(\varrho\varrho)$ au niveau de confiance $1 - \alpha$ associés aux deux échantillons.

2.1. Calculs d'intervalles de confiance de $\varrho\alpha$ et $\varrho\varrho$

Supposons que lors d'une enquête on ait obtenu les résultats suivants :

	Malades (M)	Non malades (M^c)	Total
Exposés (E)	50	50	100
Non exposés (E^c)	10	90	100
Total	60	140	200

Les données peuvent être saisies sous la forme

```
data donnees;
     input exposition$ maladie$ frequence @@;
     e m 50 e nm 50 ne m 10 ne nm 90
     ;
run;
```

Le tableau précédent peut être visualisé avec

```
proc freq data=donnees;
     weight frequence;
     tables exposition*maladie;
run;
```

On commentera les valeurs supplémentaires affichées.

Exercice 3. — (i) Procéder au calcul direct des intervalles de confiance asymptotiques de π_1, π_2, $\varrho\alpha$ et $\ln(\varrho\varrho)$, et ainsi de $\varrho\varrho$, au niveau de confiance $1 - \alpha = 0{,}95$.

(ii) Les intervalles de confiance (par defaut au niveau de confiance 95 %) des risques attribuable et relatif peuvent être calculés dans la procédure `freq` précédente en ajoutant à l'instruction `table` les options `riskdiff` et/ou `relrisk`. Retrouver ainsi les résultats

précédents. (Attention les résultats ne sont pas nécessairement faciles à lire. Ils dépendent aussi de l'ordre alphabétique du codage des différentes modalités...)

Exercice 4. — On étudie l'influence du tabagisme sur le développement du cancer du poumon. Pour cela 2000 personnes saines ont été suivies pendant vingt ans. Il y a 800 fumrurs et 1200 non fumeurs.

À la fin de l'étude 100 sujets ont développé un cancer du poumon. Parmi ceux-ci, 90 sont fumeurs et 10 non fumeurs.

(i) Quel est l'estimation du risque de cancer du poumon sur une période de 20 ans attribuable au tabagisme ? Calculer si cela est possible un intervalle de confiance au niveau 95 % de ce risque. Contient-il 0 ? Comment interprétez-vous ce résultat ?

(ii) Répondez aux même questions que précédemment en remplaçant le risque attribuable par le risque relatif. Donner la signification du risque relatif. La valeur importante obtenue pour ce risque peut-elle s'expliquer par le hasard lié à l'échantillonnage ?

Exercice 5. — L'objet de l'étude suivante est le risque de maladie coronarienne. Les sujets choisis étaient tous sains et le suivi a duré six ans.

Pour les sujets agés de 55 à 57 ans ayant une concentration de cholestérol de 8,7 mmol/l, les résultats sont les suivants

	Étudiés	Maladie coronarienne
Hommes	193	21
Femmes	227	10
Total	420	31

(i) Ces deux échantillons permettent d'estimer les différents risques pendant une durée de six ans pour une certaine population que l'on précisera.

(ii) Quelle est l'estimation du risque de maladie coronarienne attribuable au sexe masculin pour cette population, quelle est sa signification ?

(iii) Calculer, si cela est possible les intervalles de confiance de niveau 95 % des risques attribuables et relatif. Interpréter les résultats.

(iv) Le hasard lié à l'échantillonage peut-il expliquer les différences observées entre hommes et femmes ? D'autres explications sont-elles envisageables ?

3. **Intervalles de confiance du rapport de deux proportions dans les études rétrospectives**

Une étude prospective est tournée vers l'avenir. Ce type d'étude n'est pas toujours réalisable en particulier si elle implique des maladies graves. On doit alors se borner à étudier sur des groupes d'individus ayant contracté ou non par le passé s'ils ont été exposés à un certain facteur de risque. C'est ce que l'on appelle une *étude rétrospective* ou encore *étude cas-témoins*.

Pourtant, les paramètres intéressants demeurent

$$\pi_1 = \mathbf{P}(M \mid E) \qquad \text{et} \qquad \pi_2 = \mathbf{P}(M \mid E^c),$$

ainsi que les risques attribuable et relatif. Cependant, ceux-ci ne sont pas accessibles à l'estimation dans ce type d'étude. Ne peuvent être estimés que

$$\pi_1' = \mathbf{P}(E \mid M) \qquad \text{et} \qquad \pi_2' = \mathbf{P}(E \mid M^c).$$

On considère alors l'*odds-ratio* qui est le nombre

$$o\varrho = \frac{\pi_1'/(1-\pi_1')}{\pi_2'/(1-\pi_2')}\,,$$

et que l'on peut estimer dans une étude cas-témoins.

Exercice 6. — (i) Montrer que

$$o\varrho = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\,.$$

(ii) En déduire quelles relations il y a entre π_1 et π_2 en fonction de la valeur de $o\varrho$ par rapport à 1.

(iii) De quoi s'approche l'odds-ratio lorsque la maladie est rare ?

Les résultats d'une telle étude peuvent être présentés sous la forme d'un tableau

	Malades (M, les cas)	Non malades (M^c, les témoins)	Total
Exposés (E)	a	b	$n_1 = a+b$
Non exposés (E^c)	c	d	$n_2 = c+d$
Total	$a+c$	$b+d$	$n = n_1+n_2$

L'estimation de $o\varrho$ est donnée par $or = (ad)/(bc)$. Un intervalle de confiance asymptotique pour $\ln(o\varrho)$ au niveau $1-\alpha$ est donné par

$$\left[\ln(or) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\,,\ \ln(or) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right],$$

les conditions d'approximation étant généralement a, b, c, d tous supérieurs ou égaux à 10.

3.1. Calculs d'intervalles de confiance de $o\varrho$

Reprenons le tableau `donnees` de la section précédente. On obtient l'intervalle de confiance au niveau 95 % de l'odds-ratio en faisant appel à la procédure `freq` de la manière suivante :

```
proc freq data=donnees;
     weight frequence;
     tables exposition*maladie;
     exact or;
run;
```

Les options `relrisk` ou `diskdiff` sont toujours utilisables avec l'instruction `tables` mais peut-être assez peu pertinentes dans le cas d'une étude rétrospective.

Exercice 7. — (i) Calculer à l'aide de `sas` l'intervalle de confiance au niveau 95 % associé au données précédentes.

(ii) La valeur trouvée est-elle une bonne approximation du risque relatif $\varrho\varrho$? Expliquer.

Exercice 8. — On étudie le risque de thrombose veineuse profonde (TVP) en liaison avec la prise de pilule contraceptive. Il s'agit d'une étude cas-témoins. Les cas observés étaient constitués de 155 femmes agées de 15 à 49 ans ayant développé une TVP entre 1988 et 1993. Les 169 témoins étaient des amies ou des connaissance d'autres cas.

	Cas	Témoins
Prise récente de pilule	109	65
Pas de prise	46	104
Total	155	169

On sait par ailleurs que le risque annuel de TVP chez la jeune femme est d'environ 1 %.

(i) Calculer *rr* et *or*. Laquelle de ces quantité peut être utilisée pour faire une estimation du risque relatif au niveau de la population ? Donner la signification de la valeur observée *or*.

(ii) Calculer l'intervalle de confiance de $o\varrho$. Quelles conclusions peut-on faire ?

Exercice 9. — On considère une situation fictive où l'auteur (J.J. Schlesselman) veut étudier la liaison entre alcool et maladies cardio-vasculaires. Une éventuelle association et maladie cardio-vasculaires peut être due à une troisième variable à la fois liée à la maladie et au facteur étudié, comme par exemple le tabagisme. Une telle variable est appelée *facteur de confusion*. Les « observations » sont résumées dans le tableau suivant :

	Fumeurs		Non fumeurs	
	Cas	Témoins	Cas	Témoins
Prise d'alcool	65	32	10	16
Pas de prise	7	4	22	44
Total	72	36	32	60

(i) Calculer l'estimation de l'odds-ratio de la maladie par rapport au facteur d'exposition (l'alcool) et l'intervalle de confiance au niveau $\alpha = 95\,\%$. Au vu des résultats, pensez-vous que la prise d'alcool est un facteur de risque des maladies cardio-vasculaires (on admettra que ces maladies sont rares) ?

(ii) Le facteur de confusion, le tabagisme, peut expliquer les risques de maladie qui paraissent dus à l'alcool : comme chacun le sait en effet, les gens qui boivent sont souvent des fumeurs et le tabagisme est un facteur de risque des maladies cardio-vasculaires. Pour contourner le problème, on calcule l'estimation or_i de l'odds-ratio $o\varrho_i$ pour chaque modalité i du facteur de confusion. De cette façon, on obtient des groupes exposés-non exposés où l'influence du facteur de confusion est la même. Ceci peut se faire en utilisant la procédure `freq` :

```
proc freq data=truc;
     tables factconf*exposition*maladie / relrisk cmh;
run;
```

(iii) Analyser ces résultats et conclure.

Exercice 10. — S'il reste du temps, en profiter pour copier les résultats obtenus dans le futur rapport de TP.

4. **Annexe**

4.1. Abaque

L'abaque suivant a été construit pour un niveau de confiance $1-\alpha = 0{,}95$. Pour une taille d'échantillon $n \leq 25$, elle donne l'intervalle de confiance exact pour la proportion, et, pour $n > 25$, un intervalle de confiance approximatif — moins lourd à calculer — déterminé à l'aide d'une approximation normale.

`abaque.eps`

En ordonnée, on place la proportion observée p et on obtient les bornes inférieure et supérieure de l'intervalle de confiance approximatif comme les abscisses des points d'intersection de la droite horizontale $y = p$ avec les deux courbes correspondant à la taille n de l'échantillon.

4.2. Table

La table suivante donne les bornes inférieures des intervalles de confiance de niveau $1-\alpha = 0,95$ pour une proportion, où n est la taille de l'échantillon et $p = k/n$ la proportion observée. L'intervalle de confiance est $[p_{n,k}\ ,\ 1-p_{n,n-k}]$ où les $p_{n,k}$ sont les valeurs lues dans le tableau et $p_{n,0} = 0$.

n \ k	1	2	3	4	5	6	7	8	9	10
2	0,0126	0,1581								
3	0,0084	0,0943	0,2924							
4	0,0063	0,0676	0,1941	0,3976						
5	0,0050	0,0527	0,1466	0,2836	0,4782					
6	0,0042	0,0433	0,1181	0,2228	0,3588	0,5407				
7	0,0036	0,0367	0,0990	0,1841	0,2904	0,4213	0,5904			
8	0,0032	0,0318	0,0852	0,1570	0,2449	0,3491	0,4735	0,6306		
9	0,0028	0,0281	0,0749	0,1370	0,2120	0,2993	0,3999	0,5175	0,6637	
10	0,0025	0,0252	0,0667	0,1215	0,1871	0,2624	0,3475	0,4439	0,5550	0,6915
11	0,0023	0,0228	0,0602	0,1093	0,1675	0,2338	0,3079	0,3903	0,4822	0,5872
12	0,0021	0,0209	0,0549	0,0992	0,1517	0,2109	0,2767	0,3489	0,4281	0,5159
13	0,0019	0,0192	0,0504	0,0909	0,1386	0,1922	0,2513	0,3158	0,3857	0,4619
14	0,0018	0,0178	0,0466	0,0839	0,1276	0,1766	0,2304	0,2886	0,3514	0,4190
15	0,0017	0,0166	0,0433	0,0779	0,1182	0,1634	0,2127	0,2659	0,3229	0,3838
16	0,0016	0,0155	0,0405	0,0727	0,1102	0,1520	0,1975	0,2465	0,2988	0,3543
17	0,0015	0,0146	0,0380	0,0681	0,1031	0,1421	0,1844	0,2298	0,2781	0,3292
18	0,0014	0,0137	0,0358	0,0641	0,0969	0,1334	0,1730	0,2153	0,2602	0,3076
19	0,0013	0,0130	0,0338	0,0605	0,0915	0,1258	0,1629	0,2025	0,2445	0,2886
20	0,0013	0,0123	0,0321	0,0573	0,0866	0,1189	0,1539	0,1912	0,2306	0,2720

n \ k	11	12	13	14	15	16	17	18	19	20
11	0,7151									
12	0,6152	0,7353								
13	0,5455	0,6397	0,7529							
14	0,4920	0,5719	0,6613	0,7684						
15	0,4490	0,5191	0,5954	0,6805	0,7820					
16	0,4134	0,4762	0,5435	0,6165	0,6977	0,7941				
17	0,3833	0,4404	0,5010	0,5657	0,6356	0,7131	0,8049			
18	0,3575	0,4099	0,4652	0,5236	0,5858	0,6529	0,7270	0,8147		
19	0,3350	0,3836	0,4345	0,4880	0,5443	0,6042	0,6686	0,7397	0,8235	
20	0,3153	0,3605	0,4078	0,4572	0,5090	0,5634	0,6211	0,6830	0,7513	0,8316

Exemples. — Pour un échantillon de taille $n = 5$ sur lequel une proportion observée est $3/5$, on trouve pour bornes de l'intervalle de confiance $0,1466$ et $1 - 0,0527 = 0,9473$, et, pour $n = 20$ et une proportion $12/20$, on a pour bornes $0,3605$ et $1 - 0,1912 = 0,8088$.

QUELQUES TESTS USUELS

1. **tests du χ^2**

Exemple. — On cherche à savoir si la fréquence d'une maladie est liée au groupe sanguin. Sur 200 malades observés, on a obtenu la répartition suivante :

Groupe	A	B	AB	O
Effectifs	76	18	2	104

Par ailleurs, on sait que dans la population la répartition des groupes est la suivante :

Groupe	A	B	AB	O
Pourcentages	43	7	3	47

Peut-on conclure au seuil 5 % que la maladie est liée aux groupes sanguins ?

```
/* On cr\'ee une table pour les fr\'equences observ\'ees oi
et les pourcentages pi. */
data groupesanguin;
     input oi pi @@;
     cards;
     76 0.43 18 0.07 2 0.03 104 0.47
     ;
run;
/* On ajoute \‘a la table les fr\'equences attendues. */
data groupesanguin;
     set groupesanguin;
     ni=200*pi;
run;
/* On v\'erifie que les conditions d'application du test asymptotique
sont v\'erifi\'ees, par exemple ni>4. */
proc print data=groupesanguin;
run;
/* On ajoute a la table les (oi-ni)**2/ni.*/
data groupesanguin;
     set groupesanguin;
     xi=(oi-ni)**2/ni;
run;
/* On calcule la valeur de la statistique observ\'ee d= ...+xi+...
que l'on stocke dans une table test.*/
proc univariate data=groupesanguin noprint;
     var xi;
     output out=test sum=d;
run;
/* On ajoute \‘a la table la valeur seuil du test ainsi que la p-valeur.
Noter le degr\'e de libert\'e...*/
```

```
data test;
     set test;
     khi2=cinv(0.95,3);
     p=1-probchi(d,3);
run;
proc print data=test;
run;
```

Exercice 1. — Préciser le contexte de l'étude, la statistique de test utilisée et la validité du test employé. Conclure.

Évidemment, le programme Sas proposé est quelque peu « manuel ». La procédure `freq` peut effectuer ces calculs :

```
proc freq data=...;
     weight ...;
     tables ...*... / expected chisq;
run;
```

Cependant, dans ce cas, la table de données doit être correctement codée.

Exemple. — Un test au DNCB — dinitrochlorobenzène — a été fait sur des patients atteints d'un cancer de la peau et ce, à différents stades de ce cancer.

		Stades		
		I	II	III
Réaction	+	39	39	26
au DNCB	-	13	19	37

La réactivité au DNCB est-elle indépendante de la gravité du cancer de la peau ?

```
data donnees;
     input dncb$ stade$ frequence @@;
     cards;
     + I 39 + II 39 + III 26
     - I 13 - II 19 - III 37
     ;
run;

proc freq data=donnees;
     weight frequence;
     tables dncb*stade /expected chisq;
run;
```

Commenter les résultats.

Exercice 2. — Dans une étude portant sur l'orientation spatiale de la souris, les animaux de l'expérience ont été placés un par un au centre d'un labyrithe radiare comportant 8 allées arientées vers les différents points cardinaux ; chaque animal s'est échappé par l'une des allées. Les expériences ont porté, d'une part, sur des souris de laboratoire et, d'autre part, sur des souris sauvages récemment capturées en un lieu situé au nord-est du laboratoire. Les répartitions des directions de fuite sont données dans le tableau suivant :

Direction	N	NO	O	SO	S	SE	E	NE
Souris de laboratoire	17	25	13	28	19	20	22	16
Souris sauvages	26	17	9	2	3	16	33	54

Tester dans le cas des souris de laboratoire d'une part, dans le cas des souris sauvages d'autre part, l'hypothèse selon laquelle le choix de la direction se fait au hasard.

Exercice 3. — Le tableau de contingence suivant indique le résultat de l'examen sur 125 individus de leurs couleurs des yeux et des cheveux :

	Couleur des cheveux			
Couleur des yeux	blond	brun	noir	roux
bleu	25	9	3	7
Gris ou vert	13	17	10	7
Marron	7	13	8	5

Tester si il y a une dépendance entre les caractères couleur des yeux et couleur des cheveux.

2. Quelques autres tests

Il y a presque autant de tests statistiques que de situation envisageables. Nous regardons ensuite la comparaison de deux moyennes faite sur des échantillons appariés : deux mesures différentes ont été effectuées sur chaque individu et on regarde si leurs moyennes sont significativement différentes. Il y a là une difficulté supplémentaire par rapport au cas de deux échantillons indépendants : la normalité des différences n'est pas garantie.

Exemple (Un exemple fictif). — La population est l'ensemble des pièces d'un certain type et nous mesurons une certaine quantité sur celles-ci à l'aide de deux appareils. L'échantillon est constitué de 10 pièces et est supposé représentatif. Y correspond deux séries d'observations qui sont les mesures obtenues par un premier et un deuxième appareil. Les données sont donc appariées.

```
goptions reset=global gunit=pct border cback=white
colors=(blue green red yellow) ctext=red
ftitle=swissb ftext=swiss htitle=4 htext=2;

title1 "Exemple 1 : comparaison de pièces";

data exemple1;
     input app1 app2;
     diff=app1-app2;
     y=1;
     cards;
     66 61
     61 63
     63 64
     59 61
     64 64
     61 62
     60 64
     64 61
     69 63
     65 66
     ;
run;

proc print data=exemple1;
run;
```

Nous désirons comparer les moyennes des mesures obtenues par les deux appareils. Pour cela nous effectuerons un test à l'aide des différences observées. Ce test (de Student), nécessite qu'une hypothèse préalable soit satisfaite : l'hypothèse de normalité. Déja, quelques représentations graphiques :

```
symbol v=circle cv=red h=3;

proc gplot data=exemple1;
     plot y*diff;
run;

symbol2 i=boxfj co=white cv=yellow;

proc boxplot data=exemple1;
     plot diff*y;
run;
```

Puis, on utilise la procédure `univariate` :

```
proc univariate data=exemple1 noprint;
     var diff;
     histogram diff;
     output out=tests
     mean=moyenne var=variance
     probn=normal
     t=t probt=probt
     signrank=s probs=probs;
run;

proc print data=tests;
quit;
```

Pour l'hypothèse de normalité, la p-valeur obtenue est $0.15833 > \alpha = 0.05$. Nous pouvons donc conserver l'hypothèse de normalité : les observations ne sont pas en contradiction significative avec une telle hypothèse. Pour le test de Student, on a $t = 0.29032$ et p-valeur $= 0.77815 > \alpha$: les observations ne contredisent pas significativement l'hypohèse selon laquelle la moyenne de la différence puisse être nulle. Pour le test de Wilcoxon de symétrie de la distribution s $= 0.5$ (sommes des rangs des signes positifs) et une p-valeur $= 0.97656$: l'hypothèse selon laquelle la différence puisse avoir une distribution symétrique par rapport à 0 n'est pas significativement contredite par les observations.

Exercice 4 (Maladie de Hodgkin). — La population est la population des personnes atteintes de la maladie de Hodgkin (La maladie de Hodgkin est une maladie maligne du système lymphatique observée surtout chez l'adulte jeune, l'adolescent et le grand enfant. La maladie est essentiellement évoquée devant des adénopathies suspectes. Les examens radiologiques et biologiques viennent confirmer le diagnostic et permettre la classification pronostique d'Ann Arbor. Cette classification va déterminer les choix thérapeutiques.) et nous étudions la concentration dans le sang des cellules T4 et T8. Un échantillon supposé représentatif de taille 20 de malades est selectionné sur lequel les mesures de concentration en T4 et T8 sont effectuées (en nombre de cellules par mm^3)

La liste des relevés est la suivante : (396, 836), (568, 978), (1212, 1678), (171, 212), (554, 670), (1104, 1335), (257, 272), (435, 446), (295, 262), (397, 340), (288, 236), (1004, 786), (431, 311), (795, 449), (1621, 811), (1378, 686), (902, 412), (958, 286), (1283, 336), (2415, 936).

Mener une étude semblable à la précédente.

Correction. — Pour le programme :

```
proc print data=exercice1;
run;

symbol v=circle cv=red h=3;

proc gplot data=exercice1;
     plot y*diff;
run;

symbol2 i=boxfj co=white cv=yellow;

proc boxplot data=exercice1;
     plot diff*y;
run;

proc univariate data=exercice1 noprint;
     var diff;
     histogram diff;
     output out=tests
          mean=moyenne var= variance
          probn=normal
          t=t probt=probt
          signrank=s probs=probs;
run;
```

Comme d'habitude la sortie de univariate est très longue. La variable sur laquelle se fait les calculs est la variable de différence. Pour 20 observations, on a une moyenne de 209.3 (ce qui est plutot gros), un écart-type de 506.336179 (idem). La médiane est de 54.500. Elle est donc très inférieure à la moyenne ce qui laisse soupçonner une dissymétrie positive assez marquée.

```
proc print data=tests;
run;

quit;
```

Le test de normalité a pour p-valeur 0.17505 : les observations ne sont pas en contradiction significative avec l'hypothèse que les différences soient distribuées suivant une loi normale. Le test de Student est alors légitime et donne t = 1.84861 et une p-valeur 0.080139. Donc on peut accepter l'hypothèse selon laquelle la moyenne des différences est nulle. Pour le test de Wilcoxon de symétrie, s = 42, p-valeur = 0.12309. On peut donc accepter l'hypothèse de symétrie (c'est plus général que le test de normalité suivi du test de Student).

On pourra, s'il reste du temps, revenir sur des travaux pratiques précédents et/ou compléter son rapport.