2.9 Travaux pratiques

1. Dans le répertoire `TP_L3MIASHS_Stat2` qui a été créé sur le bureau.
2. Lancer `R` et modifier le répertoire de travail en allant dans `Fichier -> Changer le Répertoire Courant` et en choisissant le répertoire `Bureau/TP_L3MIASHS_Stat2` qui a été créé.
3. Ouvrir une fenêtre d'éditeur `Fichier -> Nouveau Script`.
4. Sauver le fichier dans le répertoire courant sous le nom `TP1.R : Fichier -> Sauver sous`
5. Pour les différentes questions, on peut utiliser un "copier-coller" à partir de ce document. *Il est fortement recommandé de saisir toutes les commandes dans la fenêtre ouverte de l'éditeur.* Pour exécuter les commandes saisies, il suffit de les selectionner avec la souris et d'appuyer simultanément sur les touches `Ctrl et R`.
6. Pour inclure des commentaires dans le programme, ce qui est fortement recommandé, utiliser le caractère #. Tout ce qui suit le caractère # sera négligé lors de l'exécution.
7. Penser à sauvegarder régulièrement le contenu du fichier `TP1.R` en appuyant sur les touches `Ctrl et S`.

2.9.1 Biais et Variance

La convergence d'un estimateur précise le comportement de celui-ci lorsque le nombre de données n tend vers l'infini. En estimation paramétrique, on dispose en général d'un échantillon de taille n fixée. Il convient de mettre en évidence certaines propriétés des statistiques construites sur un tel échantillon aléatoire. En fait, le point de vue est différent : "faire parler" le seul n-échantillon observé dont on dispose suppose de le considérer comme une réalisation particulière, aléatoire, d'un ensemble beaucoup plus vaste de n-échantillons possibles. Ainsi une statistique observée comme la moyenne d'échantillons $\bar{x}$ doit être vue comme la réalisation sur l'échantillon observé de la variable $\bar{X}$ qui aurait pris des valeurs différentes sur d'autres échantillons tirés au hasard.

Parmi les propriétés attendues d'un estimateur, il y a d'une part l'absence du biais de celui-ci, qui est une qualité forte, puis sa variance minimale.

Exercice 11. (Calcul de la variance empirique avec `R`)
La variance empirique d'une série d'observations $x_1, \dots, x_n$ est donnée par

$$s_n^{2\,\prime} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

1. La fonction `var()` de `R` donne-t-elle la variance empirique ? Si ce n'est pas le cas, à quelle définition est associée la fonction `var()` ?
2. Programmer une fonction qui donne la variance empirique.

La variance corrigée S_n^2 est associée (contrairement à la variance empirique) à un estimateur sans biais de la variance de population. Des illustrations très pédagogiques existent dans Wonnacot & Wonnacot (1995). Nous en reprenons la Figure 2.3 pour illustrer les définitions :

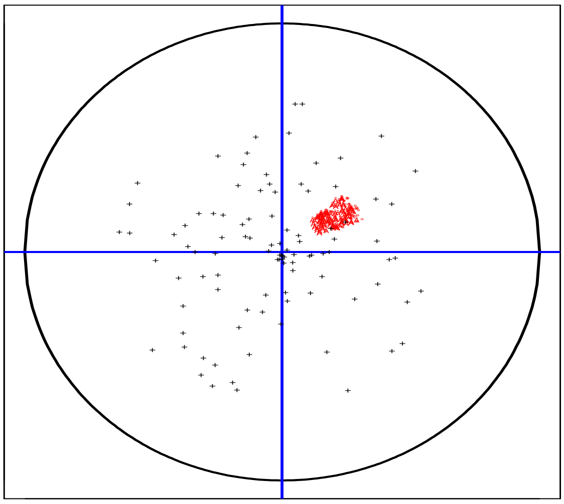


FIGURE 2.3 – Estimateur sans biais en noir, biaisé mais de faible variance en rouge

Imaginons que la valeur cible corresponde au centre du cercle. Les points noirs (resp. rouges) correspondent à 100 réalisations d'un estimateur (resp. d'un autre estimateur) de cette valeur cible. En moyenne, l'estimateur noir atteint la cible : il est sans biais. L'estimateur rouge, lui, en moyenne, se trompe : il est biaisé. En revanche sa variance est plus faible dans la mesure où les réalisations sont globalement plus proches les unes des autres.

Exercice 12. (Minimum de lois uniformes)
Soit $U_1, \ldots, U_n$ des variables aléatoires indépendantes de même loi uniforme sur $[0, 1]$. On pose $\widehat{\theta}_n = \inf(U_1, \ldots, U_n)$. $\widehat{\theta}_n$ est-il un estimateur sans biais de la borne inférieure de l'intervalle ?

Exercice 13. (Moyenne d'un échantillon)
Montrer que pour (X_n) i.i.d. selon une loi de X telle que $\mathbb{E}[X] = \mu$ et $var(X) = \sigma^2$, on a pour tout n :

$$\mathbb{E}(\bar{X}_n) = \mu \quad \text{et} \quad var(\bar{X}_n) = \frac{\sigma^2}{n} \quad \text{avec} \quad \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Dans ce cadre, la moyenne d'échantillon $\bar{X}_n$ est un estimateur sans biais de μ. Illustrer ceci pour $X \sim \mathcal{E}(\lambda)$.

Exercice 14. (Estimateur sans biais de la variance)
On a évoqué à l'Exercice 103, deux estimateurs convergents vers σ^2 :

$$S_n^{2\prime} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{et} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Vérifier que $S_n^{2\prime}$ est un estimateur sans biais de σ^2. Pour le montrer, considérer $X_i - \bar{X}_n = X_i - \mu + \mu - \bar{X}_n$.

On observe en Figure 2.3 que si l'estimateur représenté en noir est manifestement moins biaisé que le rouge, il est en revanche beaucoup plus variable puisque sa dispersion est beaucoup plus importante. Il est naturel de se demander lequel de ces deux estimateurs est finalement le plus satisfaisant : a-t-on plus de chances d'être proche de la valeur cible en tirant aléatoirement un point noir ou un point rouge ? Quelle est en moyenne l'estimateur le plus proche de la valeur cible ? Formellement, et en considérant l'erreur quadratique, il s'agit de savoir quel estimateur $\widehat{\theta}_n$ réalise le minimum de $\mathbb{E}[(\widehat{\theta}_n - \theta)^2]$ où θ est le paramètre à estimer. Cette erreur quadratique moyenne peut être écrite en fonction du biais et de la variance de l'estimateur. Plus précisement, c'est la somme du carré du biais et de la variance :

$$\mathbb{E}((\widehat{\theta}_n - \theta)^2) = (\mathbb{E}(\widehat{\theta}_n) - \theta)^2 + \mathbb{E}((\widehat{\theta}_n - \mathbb{E}(\widehat{\theta}_n))^2) = b_n^2(\widehat{\theta}_n) + var(\widehat{\theta}_n). \tag{2.9.1}$$

En pratique cette quantité est inconnue (tout comme le biais d'ailleurs) puisque θ est inconnu. Cependant, c'est souvent cette erreur qui est retenue comme mesure de la qualité d'un estimateur dans les études théoriques. Dans une étude simulée, les quatités en présence dans l'équation (2.9.1) peuvent être approchées. C'est l'objet de la section suivante.

2.9.2 Comparaison des estimateurs du paramètre de position

Soit $X_1, ..., X_n$ des variables aléatoires i.i.d. à densité. On suppose que cette densité, notée f, est symétrique par rapport à une valeur réelle θ, c'est-à-dire que

$$f(\theta + x) = f(\theta - x), \forall x \in \mathbb{R}.$$

Le but de cet exercice est d'étudier les propriétés de trois estimateurs de θ : la moyenne, la médiane et le "mid-range", définis par

$$\hat{\theta}_1 = \frac{X_1 + \cdots + X_n}{n}, \quad \hat{\theta}_2 = Med_n \quad \text{et} \quad \hat{\theta}_3 = \frac{\min_i X_i + \max_i X_i}{2}.$$

1. On commence par poser $n = 1000$ et par générer n réalisations indépendantes d'une v.a. de loi gaussienne $\mathcal{N}(5, 4)$:
```
n=1000 ;
X=randn(n,mean=5,sd=2) ;
```
2. On vérifie que la répartition des éléments de X est proche de la loi gaussienne :
```
hist(X,breaks=20,freq=F,col="cyan")
curve(dnorm(x,mean=5,sd=2),add=T)
```
Question : Quelle est la valeur de θ dans ce cas ?
3. On veut étudier les comportements des estimateurs $\hat{\theta}_1, \hat{\theta}_2$ et $\hat{\theta}_3$ lorsque $n \to \infty$.
   (a) Pour cela, on écrit la fonction suivante :
```
location_estimator=function(U,theta)
{
   n=length(U)
   theta1=cumsum(U)/(1 :n)
```

```
    theta2=1 :n
    for (i in 1 :n)
    theta2[i]=median(U[1 :i])
    end
    plot(1 :n,theta1,type="l",main="moyenne")
    abline(h=theta,col="darkred")
    plot(1 :n,theta2,type="l",main="mediane")
    abline(h=theta,col="darkred")
    return(matrix(c(theta1,theta2),n,2))
}
```

Questions :

i. Si U est un vecteur dont les coordonnées sont $(u_1, ..., u_n)$, que représente le vecteur `theta1` ? le vecteur `theta2` ?

ii. A quoi sert l'option `type="l"` dans la commande `plot` ?

(b) On appelle cette fonction à l'aide des commandes

```
X=rnorm(n,mean=5,sd=2)
est=location_estimator(X,5)
```

Question : Que remarque-t-on ? A quoi correspond la matrice `est` ?

(c) Compléter la fonction `location_estimator` pour qu'elle renvoie également la valeur de $\hat{\theta}_3$ (on pourra utiliser les commandes `cummax` et `cummin`). L'estimateur $\hat{\theta}_3$ est-il convergent ?

4. On veut maintenant comparer les 3 estimateurs.

(a) Pour cela, on écrit une fonction qui génère N échantillons dont chacun est de taille n et est distribué suivant la loi normale $\mathcal{N}(5, 4)$.
Pour chaque échantillon, on calcule les 3 estimateurs. Cela nous donne 3 séries numériques de N valeurs $\hat{\theta}_1^j$, $j = 1, \ldots, N$, $\hat{\theta}_2^j$, $j = 1, \ldots, N$ et $\hat{\theta}_3^j$, $j = 1, \ldots, N$.
On trace ensuite les boxplots de ces 3 séries numériques.

```
compare_estimators=function(N,n)
{
    X=matrix(rnorm(N*n,mean=5,sd=2),N,n)
    theta1=apply(X,1,mean)
    theta2=apply(X,1,median)
    theta3=(apply(X,1,min)+apply(X,1,max))/2
    boxplot(theta1,theta2,theta3,col="cyan")
}
```

Question : Que fait la commande `apply` ?

(b) On appelle cette fonction : `compare_estimators(200,1000)`
Au vue de ce résultat, lequel des 3 estimateurs préfère-t-on ?

(c) On modifie la fonction `compare_estimators` en remplacant la loi normale par la loi uniforme sur $[0, 10]$: `runif(N*n,0,10)`. Lequel des 3 estimateurs préfère-t-on dans ce cas-là ?

(d) On modifie encore la fonction `compare_estimators` en remplaçant la loi uniforme par la loi de Cauchy : `rcauchy(N*n,location=5,scale=2)`.

i. On observe d'abord que le troisième estimateur est incontestablement le moins bon des trois (le graphe correspondant doit être inclu dans le compte-rendu.)

ii. On supprime le calcul de $\hat{\theta}_3$ ainsi que l'affichage de son `boxplot` de la fonction `compare_estimators` et on relance la commande : `compare_estimators(200,1000)`
Le résultat obtenu est-il en faveur de $\hat{\theta}_1$ ou $\hat{\theta}_2$?
Résumer les réponses obtenues dans le tableau suivant :

Loi	$\mathcal{N}(5,4)$	$\mathcal{C}(5,2)$	$\mathcal{U}_{[0,10]}$
Meilleur estimateur parmi $\hat{\theta}_1$, $\hat{\theta}_2$ et $\hat{\theta}_3$			

(e) Si l'on dispose de vraies données, de loi inconnue, réparties de facon symétrique par rapport à une valeur θ, que fait-on pour estimer cette valeur ?

2.9.3 Intervalles de confiance (IC)

IC pour la moyenne d'un échantillon gaussien : on observe un n-échantillon, $x_1, \ldots, x_n$, distribué suivant la loi gaussien $\mathcal{N}(\mu, \sigma^2)$ où $\mu \in \mathbb{R}$ et $\sigma > 0$ sont des paramètres inconnus. Le but est de déterminer un IC de niveau 95% (par exemple) pour μ et d'étudier ses propriétés statistiques. On sait que μ et σ^2 sont bien estimées respectivement par la moyenne empirique

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

et par la variance empirique sans biais

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2.$$

1. La théorie dit que la variable aléatoire $\widehat{\theta}_n = \dfrac{(\bar{X}_n - \mu)\sqrt{n}}{S'_n}$ suit la loi de Student $t_{(n-1)}$ à $(n-1)$ degrés de liberté (cf. polycopié du cours). On cherche à avoir une confirmation empirique de ce résultat. Pour cela, on génère N échantillons de taille n de loi $\mathcal{N}(\mu, \sigma^2)$, pour des valeurs de μ et de σ^2 données. Pour chacun de ces N échantillons, on calcule la valeur de $\widehat{\theta}_n$. Cela nous donne une série numérique $t_n^1, \ldots, t_n^N$.
```
check_student=function(N,n,mu,sigma)
{
  X=matrix(rnorm(N*n,mean=mu,sd=sigma),N,n)
  t=sqrt(n)*(apply(X,1,mean)-mu)/apply(X,1,sd)
```

```
  return(t)
}
```

Si cette série est vraiment distribuées suivant la loi de Student t_{n-1}, alors son histogramme devrait être proche de la densité de la loi t_{n-1}. Pour vérifier cela, on exécute les commandes :

```
N=5000 ; n=30 ;
t=check_student(N,n,2,3)
hist(t,breaks=30,freq=F,col="cyan")
curve(dt(x,n-1),add=T,lwd=2)
```

2. *Questions* :
   (a) Déterminer les quantiles d'ordre 2.5% et 97.5% de la série numérique $t_n^1, \dots, t_n^N$. (On lira l'aide en ligne de la commande `quantile`).
   (b) Faire varier les valeurs de μ et de σ. Cela influence t-il le résultat ?
   (c) Augmenter N jusqu'à 40000 (si les capacités de la machine le permettent) et refaire l'expérience. Quelles sont les valeurs obtenues pour les deux quantiles ?
   (d) Soient a et b les valeurs obtenues dans la question précédente. Si $\widehat{\theta}_n$ est une v.a. distribuée suivant la loi t_{n-1}, quelle est la probabilité $\mathbb{P}(\widehat{\theta}_n \in [a, b])$?
   (e) Vérifier par le calcul que $\widehat{\theta}_n \in [a, b]$ équivaut à

$$\mu \in \left[\bar{X}_n - \frac{S_n}{\sqrt{n}}\, b, \bar{X}_n - \frac{S_n}{\sqrt{n}}\, a\right] \tag{2.9.2}$$

   Quel est le pourcentage des échantillons (parmi les N échantillons générés) pour lesquels (2.9.2) est satisfaite ?

3. *Application* : On mesure la force de compression d'un ciment en moulant de petits cylindres et en mesurant la pression X, mesurée en kg/cm^2, à partir de laquelle ils se cassent. Pour 10 cylindres utilisés on observe les pressions suivantes :

   19.6 19.9 20.4 19.8 20.5 21.0 18.5 19.7 18.4 19.4

   et on suppose que la variable aléatoire X obéit à une loi gaussienne. On veut déterminer un IC de niveau 95% pour la moyenne de X. Pour cela,
   (a) Rentrer les données :

$$X = c(19.6, 19.9, 20.4, 19.8, 20.5, 21.0, 18.5, 19.7, 18.4, 19.4)$$

   (b) Déterminer la valeur de n et les valeurs de a et de b correspondant à cette valeur de n.
   (c) En déduire l'IC demandé par la formule (2.9.2).

2.9.4 Calcul de l'EMV à l'aide de la commande `mle`

La méthode du MLE est une méthode pour trouver un estimateur (on ne sait pas s'il vérifira de bonnes propriétés, en particulier il sera souvent biaisé, mais c'est déjà un bon départ). Elle consiste

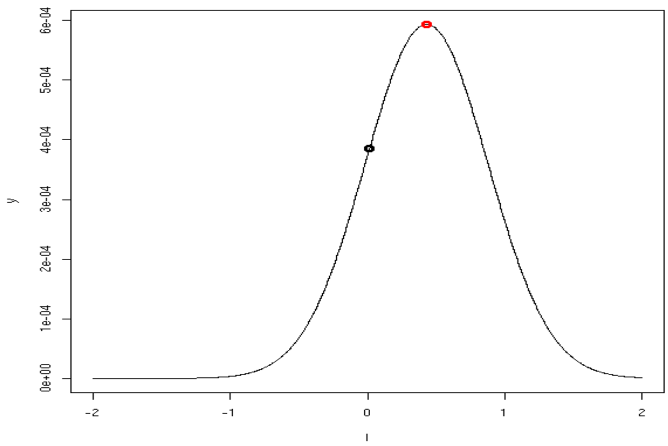


FIGURE 2.4 – Estimateur MLE

à regarder pour quelle valeur du paramètre on a le plus de chance d'obtenir le résultat qu'on a observé, i.e., à maximiser la vraisemblance.

Par exemple, cherchons, avec cette méthode, un estimateur de la moyenne d'une variable suivant une loi normale de variance 1 à l'aide d'un échantillon de 5 individus (dans cet exemple, on obtient en fait la moyenne empirique).

```
# Choix de la moyenne
m <- runif(1, min=-1, max=1)
# Les n individus
n <- 5
v <- rnorm(n, mean=m)
# Calcul de la densité
N <- 1000
l <- seq(-2,2, length=N)
y <- vector()
for (i in l) {
y <- append(y, prod(dnorm(v,mean=i)))
}
plot(y l, type='l')
# Moyenne réelle
points(m, prod(dnorm(v,mean=m)), lwd=3)
# Moyenne empirique
points(mean(v), prod(dnorm(v,mean=mean(v))), col='red', lwd=3)
```

Exercice 15. Nous souhaitons modéliser la durée de vie d'un lave-linge (en années). Pour cela nous disposons d'un échantillon de 50 valeurs :

10.273 1.035 63.952 33.153 21.708 1.612 22.216 17.682 4.730 2.236 0.746 22.762 23.455
30.993 30.138 3.140 12.884 40.893 35.560 5.999 5.962 7.069 23.743 30.881 6.647 2.393
3.736 0.866 46.587 3.728 4.173 10.127 40.502 15.275 0.030 2.517 3.134 15.530 15.156
7.919 32.617 18.881 1.221 0.360 8.641 36.070 16.217 10.520 21.354 1.366

On commence par saisir les données dans une variable qu'on appelle X :

```
X <- c(10.273, 1.035, 63.952, 33.153, 21.708, 1.612, 22.216, 17.682, 4.730, 2.236,
0.746, 22.762, 23.455, 30.993, 30.138, 3.140, 12.884, 40.893, 35.560, 5.999, 5.962,
7.069, 23.743, 30.881, 6.647, 2.393, 3.736, 0.866, 46.587, 3.728, 4.173, 10.127, 40.502,
15.275, 0.030, 2.517, 3.134, 15.530, 15.156, 7.919, 32.617, 18.881, 1.221, 0.360, 8.641,
36.070, 16.217, 10.520, 21.354, 1.366)
```

1. Représenter graphiquement cette distribution.
2. Le graphique précédent nous suggère de modéliser la durée de vie du lave-linge par une loi exponentielle $\mathcal{E}(\theta)$. On cherche maintenant estimer le paramètre θ de cette loi en utilisant la méthode du maximum de vraisemblance.
   (a) Pour cela, on définit une fonction qui calcule la log-vraisemblance négative de la loi exponentielle : `log.vrais.neg=function(theta=1){`
   ```
   if(theta>0)
   -sum(dexp(X, theta, log=TRUE))
   else NA }
   ```
   (b) Que fait la commande `dexp` appelée avec l'option `log=TRUE` ?
   (c) Faut-il minimiser ou maximiser la fonction `log.vrais.neg` pour obtenir l'EMV ?
   (d) La fonction `mle` (Maximum Likelihood Estimator) de la librairie `stats4` permet de calculer l'EMV dans un modèle paramétrique spécifié. Il faut d'abord télécharger le package `stats4` puis le charger :
   ```
   install.packages("stats4")
   library(stats4)
   fit = mle(log.vrais.neg)
   summary(fit)
   ```
   Quelle est l'estimation obtenue pour θ ?
   (e) Nous avons vu dans le cours que $\widehat{\theta}_n = 1/\bar{X}_n$ est l'EMV de θ. En déduire une estimation pour θ puis la comparer l'estimation précédente.
3. Superposer sur le graphique de la question 1 la densité de probabilité de la loi $\mathcal{E}(\widehat{\theta}_n)$.

Remarque 2.9.1. Un estimateur de maximum de vraisemblance n'est pas forcément unique (la vraisemblance peut avoir plusieurs maxima), ni sans biais, ni de variance minimale, ni efficace. Mais il possede d'excellentes propriétés asymptotiques, pour peu que la loi des observations vérifie les conditions de régularité déja évoquées pour la quantité d'information.

Théorème 2.9.1. *Si les X_i sont indépendantes et de meme loi dépendant d'un parametre réel θ, cette loi vérifiant des conditions de régularité (cf. cours correspondant), on a :*

$$\sqrt{I_n(\theta)}(\widehat{\theta}_n - \theta) \to \mathcal{N}(0,1) \text{ en loi}$$

Ceci signifie que, $\widehat{\theta}_n$ est asymptotiquement gaussien, sans biais et efficace.

Exercice 16. Illustrer par des simulations la convergence en loi de l'EMV dans chacun des cas suivants :

— un échantillon aléatoire de loi $\mathcal{E}\left(\frac{1}{\theta}\right)$ avec $\theta = 50$.

— un échantillon aléatoire de loi $\mathcal{P}(\lambda)$ avec $\lambda = 5$.

Exercice 17. *EMM versus EMV*

Nous avons vu que la méthode des moments et la méthode du maximum de vraisemblance donnent les memes résultats dans le cas de certaines lois de probabilité parmi les plus élémentaires (Poisson, exponentielle, normale, ...). En fait, dans la plupart des cas, les deux méthodes fournissent des estimateurs différents. C'est le cas par exemple de la loi uniforme.Pour le modèle $\mathcal{U}[0,\theta]$, l'estimateur obtenu par la méthode des moments est :

$$\widetilde{\theta}_n = 2\bar{X}_n$$

et l'estimateur du maximum de vraisemblance est :

$$\widehat{\theta}_n = \max\{X_1, \dots, X_n\}$$

Au vu des résultats obtenus, quelle méthode d'estimation préfère-t-on utiliser ?

3.7 Travaux pratiques

La séance de TP se fait sous environnement Windows, sauf si vous avez une nette préférence pour Linux. Pour commencer la séance :

1. Créer un répertoire `TP_L3MIASHS_Stat2` sur le bureau.
2. Lancer ensuite R et modifier le répertoire de travail en allant dans `Fichier -> Changer le Répertoire Courant` et en choisissant le répertoire `Bureau/TP_L3MIASHS_Stat2` que vous avez créé.
3. Ouvrir une fenêtre d'éditeur `Fichier -> Nouveau Script`.
4. Sauvez le fichier dans le répertoire courant sous le nom `TP2.R. Fichier -> Sauver sous`
5. Pour les différentes questions, vous pouvez utiliser un "copier-coller" à partir de ce document. Il est fortement recommandé de saisir toutes les commandes dans la fenêtre de l'éditeur que vous avez ouverte. Pour exécuter les commandes saisies, il suffit de les selectionner avec la souris et d'appuyer simultanément sur les touches `Ctrl et R`.
6. Pour inclure des commentaires dans le programme, ce qui est fortement recommandé, vous devez utiliser le caractère #. Tout ce qui suit ce caractère # sera négligé lors de l'exécution.
7. Penser à sauvegarder régulièrement le contenu du fichier `TP2.R` en appuyant sur les touches `Ctrl et S`.

3.7.1 La méthode du bootstrap

Nous verrons comment créer un outil permettant de simuler un lancer de dé. Utilisons dès à présent un tel outil pour lancer vingt dés virtuels :

```
> n <- 20
> res <- lancer.le.dé(n)
> res
```

Nous pouvons estimer la probabilité p d'obtenir un 4 par la proportion du nombre de fois où l'on a obtenu 4 dans l'échantillon ci-dessus : $\hat{p} = 0.15$.
Puisqu'un dé possède six faces, la valeur attendue est $1/6 \simeq 0.1667$. Nous constatons que ce n'est pas tout à fait la valeur que nous avons obtenue. Nous pouvons essayer de lancer de nouveau vingt dés virtuels pour voir ce qui se passe.

```
> res <- lancer.le.dé(n)
> res
```

Cette fois-ci nous estimons la probabilité p d'obtenir un 4 par $\hat{p} = 0.4$. Nous nous rendons compte que l'estimation a changé avec ce nouvel échantillon.
C'est ce que l'on appelle la *fluctuation d'échantillonnage.* Ainsi, lorsque l'on cherche à estimer un paramètre θ inconnu, les estimations obtenues varient en fonction des échantillons. Chaque nouvel échantillon observé donne lieu à une estimation différente de θ.
Si maintenant on augmente la taille de l'échantillon (par exemple à $n = 10000$), on observe une fluctuation beaucoup moins importante entre deux lancers de 10000 dés.

```
> n <- 10000
> res <- lancer.le.dé(n)
> sum(res==4)/n
[1] 0.1725
> res <- lancer.le.dé(n)
> sum(res==4)/n
[1] 0.1678
```

Nous cherchons à estimer par Monté-Carlo le biais et la variance de l'estimateur $\hat{p}$ (fréquence d'apparition de 4 pour n lancers de dés) du paramètre connu $p = \theta = 1/6$. Noter que p est le paramètre de la variable aléatoire X représentant l'apparition ou non d'un 4 lors du lancer d'un dé. La loi de X est une loi de Bernoulli de paramètre $p = 1/6$.

```
> n <- 20
> M <- 100000
> vec.theta.chap <- replicate(M,res <- lancer.le.dé(n) + theta.chap <- sum(res==4)/n)
> mean(vec.theta.chap)-1/6 # Estimation du biais.
[1] -0.00006716667
> var(vec.theta.chap) # Estimation de la variance.
[1] 0.006969351
```

Pour cet exemple, il est possible de montrer que $n\hat{\theta}(X_1, \ldots, X_n)$ est une variable aléatoire de loi binomiale $\mathcal{B}(n, p)$, d'espérance np et de variance $np(1-p)$ (avec ici $p = 1/6$). Ainsi, l'estimateur $\hat{\theta}(X_1, \ldots, X_n)$ est sans biais pour estimer θ et sa variance vaut $p(1-p)/n$.

On peut le vérifier numériquement :

```
> p <- 1/6
> p*(1-p)/n
[1] 0.006944444
```

Remarque 3.7.1. La technique du *bootstrap* permettra d'approcher le biais et la variance d'un estimateur donné, uniquement sur la base d'un seul échantillon $(X_1, \ldots, X_n)$ (comme cela est souvent le cas dans la vie réelle), et non pas sur la base de M échantillons comme cela est possible en simulation de Monté-Carlo où nous disposons d'un générateur des données.

Quelques techniques de simulation (d'une loi)

Commençons par répondre à la question "Pourquoi faire des simulations ?" :

— pour "vérifier", à l'aide de l'ordinateur, un résultat mathématique déjà connu ;
— pour "conjecturer", à l'aide de l'ordinateur, un résultat mathématique que l'on ne parvient pas à démontrer rigoureusement ;
— cela peut nous guider dans la démonstration d'un résultat mathématique difficile ;
— c'est souvent exigé lorsque l'on essaye de publier un résultat dans une revue scientifique ;

— cela permet de mieux raisonner comme un statisticien, comme nous avons déjà pu le voir dans ce chapitre.

Ceci étant dit, comment bien faire des simulations ? Il faut :
— bien cerner le (coeur du) problème ;
— décomposer le phénomène étudié en une suite d'expériences plus simples, simulables avec les outils mathématiques et informatiques dont on dispose ;
— écrire la trame d'un algorithme permettant de résoudre le problème ;
— transcrire cet algorithme dans un programme écrit dans un langage interprété comme `R` ;
— tester ce programme pour s'assurer qu'il fonctionne correctement ;
— transcrire (éventuellement) le programme en utilisant un langage de plus bas niveau (compilé) comme C/C^{++} ou Fortran.

Toute simulation nécessite d'être en mesure de simuler des variables aléatoires d'une loi donnée. Nous avons vu dans les chapitres précédents comment simuler quelques lois avec `R (rnorm(), runif(), ...)`. Lorsque la loi à simuler n'est pas implémentée dans `R`, il est possible d'utiliser l'une des méthodes présentées dans cette section.

Simuler à partir d'une autre loi

Il existe parfois des formules simples exprimant la variable aléatoire X de loi L, dont on veut simuler des observations, en fonction d'une ou de plusieurs variables aléatoires de lois usuelles. Il devient alors aisé de fabriquer un générateur d'observations de loi L au moyen de cette formule. L'exemple très simple suivant permet d'illustrer ce point. On se rappellera par exemple que l'on peut fabriquer une variable aléatoire de loi χ^2_1 en prenant le carré d'une variable aléatoire gaussienne standard.

`> rkhi2.1 <- function() rnorm(1)^2 # Si` $X \sim \mathcal{N}(0,1)$ `alors` $X^2 \sim \chi^2_1$.

Exercice 24. Générez des observations suivant une loi $TU(2)$ qui est centrée et dont la densité de probabilité $\dfrac{1}{F_\lambda(x)^{\lambda-1} + (1-F_\lambda(x))^{\lambda-1}}$, $X = U^\lambda - (1-U)^\lambda$ et de variance $\dfrac{2}{2\lambda+1} - 2\dfrac{\Gamma^2(\lambda+1)}{\Gamma(2\lambda+2)}$.

Méthode de la transformation inverse

Supposons que l'on connaisse la fonction de répartition réciproque F^{-1} de la variable aléatoire X et que l'on veuille simuler des observations ayant la même loi que celle de X. Cela peut alors très facilement se faire à partir d'un générateur U de loi $\mathcal{U}(0,1)$, au moyen de la formule suivante :

$$\tilde{X} = F^{-1}(U)$$

En effet, il se trouve que la variable aléatoire $\tilde{X}$ a pour fonction de répartition F_X. Cette propriété est connue sous le nom de *transformation intégrale de probabilité* et a été découverte par R.A. Fisher.

On rappelle que la fonction de répartition d'une variable aléatoire X de loi exponentielle $Exp(\lambda)$ est donnée par :

$$F_X(x,\lambda) = 1 - e^{-\lambda x} \text{ si } x \geq 0 \text{ et } 0 \text{ sinon}$$

Calculer F_X^{-1} puis utiliser la fonction `runif()` pour générer des observations d'une loi $Exp(2)$.

Exercice 25. Nous allons utiliser la méthode du rejet pour générer des observations de loi $\mathcal{N}(0,1)$ en prenant comme fonction de référence la densité d'une loi exponentielle de paramètre $\lambda = 1$. Prendre la valeur $c = \sqrt{e^1/(2\pi)}$ et utiliser la fonction `rbinom()` pour affecter un signe positif ou négatif aux valeurs fabriquées par l'algorithme du rejet.

Simulation de variables aléatoires discrètes

On veut simuler un échantillon provenant d'une variable aléatoire discrète X satisfaisant $\mathbb{P}(X = x_i) = p_i$ pour tout i dans $\mathbb{N}$ (ou l'un de ses sous-ensembles). On définit U comme étant une variable aléatoire de loi $\mathcal{U}(0,1)$ qui pourra être obtenue à l'aide de la fonction `runif()`, et on utilise l'algorithme suivant :

$$\begin{cases} X = x_0 & \text{si } 0 < U \leq p_0; \\ X = x_i & \text{si } \sum_{j=0}^{i-1} p_j < U \leq \sum_{j=0}^{i} p_j. \end{cases}$$

Exercice 26. Générer des observations de loi uniforme discrète sur $\{0,1,2,3,4,5\}$ en utilisant cet algorithme.

Reprenons l'exemple où nous cherchions à estimer le biais et la variance de l'estimateur $\hat{p}$ (fréquence d'apparition du chiffre 4 pour n lancers de dés) du paramètre $p = \theta$ de la variable aléatoire X représentant l'apparition ou non d'un 4 lors du lancer d'un dé.

```
> n <- 20
xvec <- lancer.le.de(n)
sum(xvec==4)/n
[1] 0.15
# Tirons avec remise un éch. de taille n dans xvec.
> sample(xvec,n,replace=TRUE)
[1] 6 6 4 5 6 2 5 2 5 2 5 4 6 6 5 5 6 2 1 4
> B <- 10000
> vec.theta.etoile <- replicate(B,sum(sample(xvec,n,replace=TRUE)==4)/n)
> mean(vec.theta.etoile) - sum(xvec==4)/n
[1] 0.00009
> ((B-1)/B)*var(vec.theta.etoile)
sd(vec.theta.etoile)
[1] 0.006377992
[1] 0.07986632
```

Dans ce cas très simple, la théorie nous dit que le biais est nul et que la variance vaut $p(1-p)/n = 0.00694$.

Astuce : noter l'existence du package `boot` qui facilite la pratique du bootstrap :

`boot(xvec,function(x,w) sum(x[w]==4)/n,B)`

Exercice 27.

A. Etude de la loi $f(x) = \frac{3}{2}\sqrt{x}$ sur $[0, 1]$

(a) Vérifier que $f(x)$ est une densité au moyen de la fonction `integrate()`.

(b) Simuler un échantillon de taille 1000 selon la loi définie par la densité $f(x) = \frac{3}{2}\sqrt{x}$ sur $[0, 1]$

(c) Calculer les moyennes et variances empiriques.

(d) Comparer avec les valeurs théoriques.

(e) Calculer et comparer les probabilités théoriques et empiriques des classes suivantes :

$$[0, 0.30],\ ]0.30, 0.50],\ ]0.50, 0.70],\ ]0.70, 0.85],\ ]0.85, 1].$$

B. Théorème de Box et Muller.
Soit U_1 et U_2 deux variables aléatoires uniformes et indépendantes sur l'intervalle $[0, 1]$. Les variables

$$Z_1 = \sqrt{-2\log(U1)}\cos(2\pi U_2) \text{ et } Z_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

sont alors deux variables aléatoires normales centrées réduites indépendantes.

(a) Générer $n = 1000$ couples d'observations $(z_1, z_2)_1, \ldots, (z_1, z_2)_n$ en utilisant cet algorithme.

(b) Utiliser la fonction `kde2d()` du *package* `MASS` pour estimer la densité bivariée de ces données.

(c) Utiliser les fonctions `spheres3d()` et `surface3d()` du *package* `rgl` pour représenter ces observations, ainsi que la surface estimée de la densité bivariée de ces données et également la surface de la densité d'une gaussienne standard bivariée.

(d) Constater que l'on obtient une courbe en cloche caractéristique de la gaussienne bivariée.

3.7.2 Estimation fonctionnelle

Exercice 28.

1. Générer 200 observations de loi lognormale $L\mathcal{N}(0, 1)$ (aide : générer 200 observations de loi $\mathcal{N}(0, 1)$ et les transformer par e^x).
2. Donner une estimation ponctuelle et par intervalle pour la médiane de la loi.
3. L'intervalle contient-il la vraie valeur ? Recommencer pour estimer le quantile d'ordre 0.90.

Exercice 29. (Adapté de Mosteller et Tukey, 1977)
Soit les valeurs

$$0.1\ ,\ 0.1\ ,\ 0.1\ ,\ 0.1\ ,\ 0.4\ ,\ 0.5\ ,\ 1.0\ ,\ 1.1\ ,\ 1.3\ ,\ 1.9\ ,\ 1.9\ ,\ 4.7$$

1. Donner une estimation du jackknife de l'écart-type de la loi mère ayant généré ces observations, fondée sur la statistique S_n.
2. Donner un intervalle de confiance pour cet écart-type.

Exercice 30.

1. Générer 50 observations de la loi $\mathcal{N}(0,1)$.
2. Estimer $f(0)$ par un noyau biweight avec $h = 1$.
3. Comparer à la vraie valeur.
4. Recommencer la procédure plusieurs fois pour confirmer le type de biais en présence.

Exercice 31.

1. Générer 50 observations de loi $\mathcal{N}(0,1)$.
2. Estimer $f(0)$ par un noyau biweight avec $h = 0.5, 0.75, 1, 1.25, 1.5$ pour apprécier la variabilité des estimations ainsi obtenues.
3. Quelle est la valeur de h asymptotiquement optimale ?

Estimation à noyau de la densité de probabilité

Rappelons l'estimateur à noyau de la densité de probabilité :

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

— Tout estimateur à noyau est composé d'une somme de n fonctions, soit une pour chaque observation.

```
> x = c( -1.111, -0.257, 1.797, 2.163, -2.2264, -0.949)
> K = function(x,a) dnorm(x,a,1.09)/6 # noyau gaussien, n = 6, h = 1.09
(règle de Silverman)
> b = seq(-6,6,.01)  # valeurs où l'on calcule K
> for (i in 1 :6) { plot(b,K(b,x[i]),type=l,ylab= , xlim=c(-7,7),
                  ylim=c(0,.06),col=blue) ; par(new=T)}
#chaque observation détermine un graphique
#xlim, ylim permettent une superposition parfaite des graphiques
> title(Estimateur du noyau decomposé)
```

— La forme générale de la fonction `density` est la suivante :

```
density(x, bw = "nrd0", adjust = 1, kernel = c("gaussian", "epanechnikov",
"rectangular", "triangular", "biweight", "cosine", "optcosine"), weights = NULL,
window = kernel, width, give.Rkern = FALSE, n = 512, from, to, cut = 3, na.rm
= FALSE, ...)
```

où

— `x` #vecteur de données univariées
— `bw` #paramètre de lissage
— `bw="nrd0"` # règle de Silverman
— `bw="nrd"` # règle de Scott
— `bw="ucv"` # règle de la validation croisée

— `bw="SJ-dpi"` # règle de Sheather-Jones
— `kernel` # type de noyau, par défaut “gaussian”
— n=512 # n. de points équidistants où f est estimée

Liste de longueur 7 dont les principales composantes sont :

— `[[1]]x` # vecteur des points où f est estimée
— `[[2]]y` # vecteur des valeurs de f estimées
Par défaut $n = 512$ points ou valeurs. Si $n > 512$, le n. de points ou de valeurs est la puissance de 2 supérieure à 512.
— `[[3]]bw` # paramètre de lissage h utilisé
— `[[4]]n` # n. de points où $\hat{f}$ est calculé (puissance de 2)
— L’application de density ne produit aucun graphique. On obtient plutôt de l’information sur les valeurs estimées et sur h.
— La fonction `plot` appliquée à l’objet créé par `density` produit le graphique.

Exercice 32. Pour une application de la méthode du noyau, on considère les données financières (Jeu des taux d’épargne) : `cdrate.dat`.

1. Partager, en deux, la fenêtre des graphiques et télécharger les données résiduelles :
```
> par(mfrow=c(1,2))
> taux = scan(cdrate.dat)
```
2. Donner le paramètre de lissage par défaut (Silverman, Scott, Sheather-Jones).
3. Tracer l’estimateur de la densité de probabilité pour un noyau rectangulaire puis gaussien. On utilise ici le paramètre de lissage par défaut (règle de Silverman).

Package `sm`

— Fonction `sm.density` :
`> library(sm)`
`sm.density(x, h, model = "none", weights = NA, group=NA, ...)` # estimation avec le noyau *gaussien*
— Produit le graphique de l’estimateur à noyau *gaussien*
— `x` # données uni- (vecteur), bi- ou tridimensionnelles (matrice, tableau) ;
`h` # paramètre de lissage (par défaut, règle de Scott)
— Autres arguments intéressants (voir `sm.options`)
`add=TRUE` # ajoute à un graphique la densité estimée
`display=se` # produit une bande de variabilité
`method` # choix de h, =”normal” (Scott), ”cv” (validation croisée), ”sj” (Sheather-Jones)
`hmult` # multiple du h, = 1 par défaut

Exercice 33.

1. Télécharger la librairie `sm` et le jeu de données `aircraft`. Il s’agit de 709 modèles d’avion du 20e siècle, tableau à 8 variables, incluant Span et Period. Appliquer la commande, suivante,

propre a `sm` :

```
> provide.data(aircraft)
> y = log(Span[Period==3]) #var. Span 19561984
> par(mfrow=c(2,2)) # pour mettre plusieurs graphiques sur la même page
```

2. Lancer `sm.density` en variant le paramètre de lissage $hmult = 1/3, 2, 6$:
   `sm.density(y,hmult=1/3,xlab=log(envergure)\n hm = 1/36, ylab=densité) # hmult facteur multipliant h optimal`
3. Tracer les densités de l'envergure des 709 modèles d'avion
4. Comparer les troix densités :

```
> y1 = log(Span)[Period==1] #début du 20e siècle
> y2 = log(Span)[Period==2] #milieu du 20e siècle
> y3 = log(Span)[Period==3] #fin du 20e siècle
```

5. Supperposer les trois graphiques en utilisant `add = TRUE`. On peut aussi ajouter une légende :
   `> legend(3.15,1,c(1914-1935,1936-1955,1956- 1984),lty=3 :1) # 3 périodes du 20e siècle couvertes par le jeu aircraft`
6. Construction avec `sm.density` et l'argument `display` de la bande de confiance :

```
> provide.data(aircraft)
> y = log(Span[Period==3])
> sm.density(y,xlab=log(envergure),ylab=densité, display=se)
```

— *Loi de $\widehat{P}_n$ sous H_0* : on vérifie les conditions d'application du test

$$n = 300 \geq 30,\ np_0 = 300 \times 0.75 = 225 \geq 5 \text{ et } n(1 - p_0) = 300 \times (1 - 0.75) = 75 \geq 5$$

(effectifs théoriques des deux catégories de patients, attendus sous H_0)

— La statistique $\widehat{P}_n$ suit donc approximativement la loi normale :

$$F_n \simeq \mathcal{N}\left(0.75, \sqrt{\frac{0.75 \times 0.25}{300}}\right) \text{ et } Z_n = \frac{\widehat{P}_n - 0.75}{\sqrt{\frac{0.75 \times 0.25}{300}}} \simeq \mathcal{N}(0, 1)$$

— *Méthode 1* : IA (Intervalle d'Acceptation) et RC (Région Critique) pour $\widehat{P}_n$, associés au risque $\alpha = 5\%$.
On rejette H_0 pour les valeurs de $\widehat{P}_n$ qui sont trop grandes par rapport à 0.75.
RC à l'extrémité droite du domaine de $\widehat{P}_n$ et

$$IA = \left[0,\quad 0.75 + 1.645\sqrt{\frac{0.75 \times 0.25}{300}}\right] = [0, 0.791]$$

avec 1.645 quantile d'ordre 0.95 de la loi $\mathcal{N}(0, 1)$.
Décision : $\widehat{p} = 0.81 \notin IA$, donc on rejette H_0 pour un risque d'erreur de 5%.

— *Méthode 2* : IA et RC pour Z_n, associés au risque $\alpha = 5\%$

$$IA =]-\infty, 1.645] \text{ et } z_{obs} = \frac{0.81 - 0.75}{\sqrt{\frac{0.75 \times 0.25}{300}}} = 2.4 \notin IA$$

— On peut également calculer la p-valeur α_{obs}

$$\alpha_{obs} = \mathbb{P}_{H_0}(\widehat{P}_n \geq 0.81) = \mathbb{P}_{H_0}(Z_n \geq 2.4) = 1 - F(2.4) = 0.82\%$$

Comme $\alpha > \alpha_{obs}$, on peut rejeter H_0.

4.8 Travaux dirigés

Exercice 34. Une machine produit des pièces dont une proportion θ présente des défauts qui conduisent classer ces pièces en second choix. En cas de bon fonctionnement la valeur est $\theta = 0.1$. Si la machine se dérègle, la valeur passe $\theta = 0.2$. On estime que la probabilité que la machine soit bien réglée est $p_0 = 0.6$. On peut souscrire un contrat d'entretien qui permet de maintenir la machine bien réglée. Le coût est de 0.4 euros par pièce produite. Chaque pièce de second choix est vendue 6 euros de moins que les autres. Après contôle des 100 premières pièces produites, on constate que 13 d'entre elles doivent être classées en second choix. Quelle décision est-on conduit prendre si on utilise la méthode de Bayes ?

Exercice 35. Pour savoir si un enfant est ambidextre, on lui tend un objet 12 fois, et on note à chaque fois de quelle main il se sert pour le saisir. Soit X le nombre de fois où il se sert de sa main droite.

1. L'hypothèse nulle est que l'enfant est ambidextre.
   (a) Formuler cette hypothèse en fonction d'un paramètre p.
   (b) Expliquer clairement la signification de p.
   (c) Formuler une alternative H_1.
   (d) Déterminer un test à 5%.
2. Si l'hypothèse nulle est vraie, quelle est la probabilité de la rejeter ?
3. Pourquoi n'est-on pas capable de calculer la probabilité d'accepter H_0 lorsque H_0 est fausse ?
4. Calculer la probabilité $\phi(p)$ de rejeter H_0 lorsque $p = 0.05, 0.10, \ldots, 0.95$.
5. Tracer la fonction $\phi(p)$.

Exercice 36. Dans chacun des cas suivants, on veut tester :

$$H_0 \;:\; \theta = \theta_0 \;\text{ contre l'alternative }\; H_1 \;:\; \theta > \theta_0$$

Dire laquelle des deux régions critiques suivantes est la meilleure :

$$\mathcal{C}_1 = \{x \in \mathbb{R} \;:\; x \geq a\} \quad \text{et} \quad \mathcal{C}_2 = \{x \in \mathbb{R} \;:\; x \leq b\}.$$

1. Fonction de densité : $f(x) = \theta e^{-x\theta},\ x > 0,\ \theta > 0$.
2. Fonction de masse : $p(x) = \theta(1-\theta)^x,\ 0 < \theta < 1,\ x = 0, 1, \ldots$
3. fonction de masse : $p(x) = \theta^x e^{-\theta}/x!,\ \theta > 0,\ x = 0, 1, \ldots$
4. Fonction de densité : $f(x) = xe^{-x/\theta}/\theta^2,\ x > 0,\ \theta > 0$.
5. Fonction de densité : $f(x) = \theta c^\theta x^{-(\theta+1)},\ x > c > 0$, $\theta > 1$ où c est une constante connue.

Exercice 37. On dispose d'un échantillon $X_1, \ldots, X_n$ d'un caractère X qui suit une loi normale d'espérance inconnue μ et d'écart-type connu $\sigma = 1$ pour choisir entre les deux hypothèses :

$$\begin{cases} H_0 & : \mu = 1 \\ H_1 & : \mu = 1.5 \end{cases}$$

1. Déterminer la région critique du test de Neyman et Pearson.
2. Calculer sa puissance dans le cas où $n = 25$ et $\alpha = 0.05$.
3. Quelle devrait être la taille d'échantillon minimum pour que cette puissance soit supérieure 0.90 ?

Exercice 38. On dispose d'un échantillon $X_1, \ldots, X_n$ d'un caractère X qui suit une loi normale centrée, d'écart-type inconnu σ pour choisir entre les deux hypothèses :

$$\begin{cases} H_0 & : \sigma = 1 \\ H_1 & : \sigma = 2 \end{cases}$$

1. Déterminer la région critique du test de Neyman et Pearson.
2. Calculer sa puissance dans le cas où $n = 15$ et $\alpha = 0.05$.

Exercice 39. Le revenu annuel des individus d'une population est distribué selon une loi de Pareto de densité de probabilité :

$$f(x, \theta) = \frac{\theta}{x^{1+\theta}} \text{ si } x \geq 1 \text{ et } 0 \text{ sinon}$$

On dispose d'un échantillon $X_1, \ldots, X_n$ de cette loi pour choisir entre les deux hypothèses :

$$\begin{cases} H_0 & : \theta = 4/3 \\ H_1 & : \theta = 8/5 \end{cases}$$

1. Déterminer la région critique du test de Neyman et Pearson.
2. Calculer sa puissance dans le cas où $n = 400$ et $\alpha = 0.05$.

Exercice 40. On dispose d'un échantillon de taille $n = 15$ d'une v.a. de loi normale centrée et de variance $1/\theta$ pour choisir entre les deux hypothèses :

$$\begin{cases} H_0 & : \theta = 1 \\ H_1 & : \theta > 1 \end{cases}$$

1. Déterminer la région critique d'un test UPP de risque de première espèce α.
2. Préciser sa fonction de puissance.
3. Calculer cette puissance dans le cas où $n = 15$ pour $\theta = 3$ et $\alpha = 0.05$.

Exercice 41. Pour tester l'hypothèse

$$H_0 : \mu \leq 180 \quad \text{contre l'alternative} \quad H_1 : \mu > 180$$

où μ est la moyenne d'une population normale d'écart-type $\sigma = 24$, on prélève un échantillon de taille $n = 64$ et on décide de rejeter H_0 si $\bar{X}_{64} > 186$.

1. Quelle est la probabilité de rejeter H_0 si $\mu = 175$?
2. On ne peut pas calculer la probabilité exacte d'une erreur de première espèce. Mais, quelle est la probabilité maximale d'une erreur de première espèce ?
3. Quelle est la probabilité maximale d'une erreur de seconde espèce ?

4.9 Travaux pratiques

La séance de TP se fait sous l'environnement Windows, sauf si l'on a une nette préférence pour Linux. Pour commencer la séance :

1. Créer un répertoire `TP_L3MIASHS_Stat2` sur le bureau.
2. Lancer ensuite `R` et modifier le répertoire de travail en allant dans `Fichier -> Changer le Répertoire Courant` et en choisissant le répertoire `Bureau/TP_L3MIASHS_Stat2` qui est déjà créé.
3. Ouvrir une fenêtre d'éditeur `Fichier -> Nouveau Script`.
4. Sauver le fichier dans le répertoire courant, par exemple, sous le nom `TP3.R : Fichier -> Sauver sous`
5. Pour les différentes questions, vous pouvez utiliser un "copier-coller" à partir de ce document. *Il est fortement recommandé de saisir toutes les commandes dans la fenêtre de l'éditeur que l'on a ouverte*. Pour exécuter les commandes saisies, il suffit de les selectionner avec la souris et d'appuyer simultanément sur les touches `Ctrl et R`.
6. Pour inclure des commentaires dans le programme, ce qui est fortement recommandé, on doit utiliser le caractère #. Tout ce qui suit le caractère # sera négligé lors de l'exécution.
7. Penser à sauvegarder régulièrement le contenu du fichier `TP3.R` en appuyant sur les touches `Ctrl et S`.

4.9.1 Rappels des bases des tests statistiques

Nous rappelons, ci-dessous, quelques définitions et vocabulaires.

- *Hypothèses*. On oppose deux hypothèses complémentaires H_0 et H_1 :
  — l'hypothèse H_0 formule ce que l'on souhaite rejeter/réfuter,
  — l'hypothèse H_1 formule ce que l'on souhaite montrer.
  Par exemple, si on veut montrer l'hypothèse "lôt non conforme", H_0 et H_1 s'opposent sous la forme :
  $$H_0 : \text{"lôt conforme"} \quad \text{contre } H_1 : \text{"lôt non conforme"}.$$
- *Risque* . Le risque est le pourcentage de chances de rejeter H_0, donc d'accepter H_1, alors que H_0 est vraie. On veut que ce risque soit aussi faible que possible. Il s'écrit sous la forme : $100\alpha\%$, avec $\alpha \in]0,1[$ (par exemple, 5%, soit $\alpha = 0.05$). Le réel α est alors la probabilité de rejeter H_0 alors que H_0 est vraie. Le rejet de H_0 est dit "significatif" si elle est rejetée au risque 5%.
- *Test statistique*. Un test statistique est une procédure qui vise à apporter une réponse à la question : Est-ce que les données nous permettent de rejeter H_0, donc d'accepter H_1, avec un faible risque de se tromper ?
- *Types de test statistique*. En notant θ un paramètre inconnu, on dit que le test est :
  — bilatéral si H_1 est de la forme $H_1 : \theta \neq \cdots$
  — unilatéral à gauche (sens de <) si H_1 est de la forme $H_1 : \theta < \cdots$

— unilatéral à droite (sens de $>$) si H_1 est de la forme $H_1 : \theta > \cdots$

- *p-valeur*. La p-valeur est le plus petit réel $\alpha \in]0, 1[$ calculé à partir des données tel que l'on puisse se permettre de rejeter H_0 au risque $100\alpha\%$. Autrement écrit, la p-valeur est une estimation ponctuelle de la probabilité critique de se tromper en rejetant H_0 alors que H_0 est vraie.
- *Degré de significativité*. La p-valeur nous donne un degré de significativité du rejet de H_0. Le rejet de H_0 sera :
  — significatif si p-valeur $\in]0.01, 0.05]$, symbolisé par *,
  — très significatif si p-valeur $\in]0.001, 0.01]$, symbolisé par **,
  — hautement significatif si p-valeur < 0.001, symbolisé par ***.
  Il y a non rejet de H_0 si p-valeur > 0.05.
- *En cas de non-rejet de H_0*. S'il y a non-rejet de H_0, sauf convention, on ne peut rien conclure du tout (avec le risque considéré). En revanche, peut-être qu'un risque de départ plus élevé ou la disposition de plus de données peuvent conduire à un rejet de H_0.

4.9.2 Tests de conformité à une valeur de référence

— *Enjeu*. L'enjeu d'un test de conformité est d'affirmer, avec un faible risque de se tromper, qu'une norme associée à un caractère X (sa moyenne, une proportion, ...) n'est plus conforme à la réalité. Ainsi, en posant H_1 : "la norme n'est plus conforme", on se pose la question : Est-ce que les données $x_1, \ldots, x_n$, qui sont des observations de X, nous permettent de rejeter H_0, donc d'accepter H_1, avec un faible risque de se tromper ?

— *Formules : p-valeurs* (cf. Table 6.4).
Lois :

$$Z \sim \mathcal{N}(0,1),\ T \sim t_\nu \text{ et } K \sim \chi^2_\nu \text{ où } \nu = n-1$$

Outils :

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

— *Commandes*. Pour les commandes (cf. Table 6.5), on considère les librairies `stats` et `OneTwoSamples` :

```
library(stats)
library(OneTwoSamples)
```

$X \sim \mathcal{N}(\mu, \sigma^2)$	H_1	Stat. test obs.	p-valeurs
σ connu `Z-Test`	$\mu \neq \mu_0$ $\mu > \mu_0$ $\mu < \mu_0$	$z_{obs} = \sqrt{n}\left(\frac{\bar{x} - \mu_0}{\sigma}\right)$	$\mathbb{P}\left(\lvert Z\rvert \geq \lvert z_{obs}\rvert\right)$ $\mathbb{P}\left(Z \geq z_{obs}\right)$ $\mathbb{P}\left(Z \leq z_{obs}\right)$
σ inconnu `T-Test`	$\mu \neq \mu_0$ $\mu > \mu_0$ $\mu < \mu_0$	$t_{obs} = \sqrt{n}\left(\frac{\bar{x} - \mu_0}{s_n}\right)$	$\mathbb{P}\left(\lvert T\rvert \geq \lvert t_{obs}\rvert\right)$ $\mathbb{P}\left(T \geq t_{obs}\right)$ $\mathbb{P}\left(T \leq t_{obs}\right)$
`1-Chi2-Test`	$\sigma^2 \neq \sigma_0^2$ $\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$	$\chi^2_{obs} = \frac{n-1}{\sigma_0^2}\, s_n^2$	$2\min\left(\mathbb{P}\left(K \geq \chi^2_{obs}\right), \mathbb{P}\left(K \leq \chi^2_{obs}\right)\right)$ $\mathbb{P}\left(K \geq \chi^2_{obs}\right)$ $\mathbb{P}\left(K \leq \chi^2_{obs}\right)$
$X \sim \mathcal{B}(p)$	H_1	Stat. test obs. et var.	p-valeurs
$n \geq 31$, $np_0 \geq 5$ $n(1-p_0) \geq 5$ `1-Prop-Z-Test cor`	$p \neq p_0$ $p > p_0$ $p < p_0$	$z_{obs} = \sqrt{n}\left(\frac{\lvert\bar{x} - p_0\rvert - \frac{0.5}{n}}{\sqrt{p_0(1-p_0)}}\right)$	$\mathbb{P}\left(\lvert Z\rvert \geq \lvert z_{obs}\rvert\right)$ $\mathbb{P}\left(Z \geq z_{obs}\right)$ $\mathbb{P}\left(Z \leq z_{obs}\right)$
$n \geq 31$, $np_0 \geq 5$ $n(1-p_0) \geq 5$ `1-Prop-Z-Test`	$p \neq p_0$ $p > p_0$ $p < p_0$	$\sqrt{n}\left(\frac{\bar{x} - p_0}{\sqrt{p_0(1-p_0)}}\right)$	$\mathbb{P}\left(\lvert Z\rvert \geq \lvert z_{obs}\rvert\right)$ $\mathbb{P}\left(Z \geq z_{obs}\right)$ $\mathbb{P}\left(Z \leq z_{obs}\right)$

TABLE 4.9 – Tests de conformité à une valeur de référence

$X \sim \mathcal{N}(\mu, \sigma^2)$	H_1	Commandes

σ connu
Z-Test
	$\mu \neq \mu_0$
	$\mu > \mu_0$
	$\mu < \mu_0$
		`mean_test1(x, mu0, sigma)$p_value`
		`mean_test1(x, mu0, sigma, side = 1)$p_value`
		`mean_test1(x, mu0, sigma, side = -1)$p_value`

σ inconnu
T-Test
	$\mu \neq \mu_0$
	$\mu > \mu_0$
	$\mu < \mu_0$
		`t.test(x, mu = mu0)$p.value`
		`t.test(x, mu = mu0, alternative = "greater")$p.value`
		`t.test(x, mu = mu0, alternative = "less")$p.value`

1-Chi2-Test	$\sigma^2 \neq \sigma_0^2$
	$\sigma^2 > \sigma_0^2$
	$\sigma^2 < \sigma_0^2$
		`var_test1(x, sigma20)$P_value`
		`var_test1(x, sigma20, side = 1)$P_value`
		`var_test1(x, sigma20, side = -1)$P_value`

$X \sim \mathcal{B}(p)$	H_1	Commandes

$n \geq 31$, $np_0 \geq 5$
$n(1-p_0) \geq 5$
1-Prop-Z-Test cor
	$\mu \neq \mu_0$
	$p > p_0$
	$p < p_0$
		`prop.test(x, n, p)$p.value`
		`prop.test(x, n, p, alt = "greater")$p.value`
		`prop.test(x, n, p, alternative = "less")$p.value`

$n \geq 31$, $np_0 \geq 5$
$n(1-p_0) \geq 5$
1-Prop-Z-Test
	$p \neq p_0$
	$p > p_0$
	$p < p_0$
		`prop.test(x, n, p, correct = F)$p.value`
		`prop.test(x, n, p, alternative = "greater", correct = F)$p.value`
		`prop.test(x, n, p, alternative = "less", correct = F)$p.value`

TABLE 4.10 – Commandes R pour les tests de conformité à une valeur de référence

Remarque 4.9.1. En omettant les commandes `$p.value` (ou `$p_value`), les commandes renvoient plus d'éléments associés au test statistique considéré, dont la p-valeur (statistique de test observée, degré de liberté, intervalle de confiance, ...).

4.9.3 Comparaison de la moyenne théorique à une valeur de référence

- *Descriptif du test.* Soit une variable quantitative X de moyenne théorique μ et de variance σ^2. A partir d'un échantillon de taille n, on veut comparer la moyenne théorique μ à une valeur de référence μ_0. Les hypothèses du test sont $H_0 : \mu = \mu_0$ et l'hypothèse alternative peut être $H_1 : \mu > \mu_0$, $H_1 : \mu \neq \mu_0$ ou $H_1 : \mu < \mu_0$.
  Sous H_0, la statistique de test est :
  $$T_n = \sqrt{n}\left(\frac{\bar{X}_n - \mu_0}{S_n}\right) \sim t_{n-1}$$
- *Conditions de validité.* Normalité des données ou taille d'échantillon assez grande ($n > 30$).
- *Instruction R* : Il est possible d'utiliser la fonction `t.test()`.

Exemple 4.9.1. Dans le but de la mise en oeuvre de ce test, nous considérons le jeu de données Intima-Media. On veut savoir si les personnes qui ont un indice de masse corporelle (IMC) supérieur à 30 ont une mesure de l'épaisseur de l'intima-média supérieure en moyenne à la mesure dans la population dont est issu l'échantillon.
On suppose que cette moyenne théorique de la mesure intima-média dans cette population est égale à 0.58 mm.

```
IMC <- poids/((taille/100)^2)
mesurel <- mesure[IMC>30]
t.test(mesurel,mu=0.58,alternative="greater")
```

4.9.4 Comparaison de la variance théorique à une valeur de référence

- *Descriptif du test.* Soit σ^2 la variance d'un caractère quantitatif X de la population $\mathcal{P}$. Les hypothèses du test sont $H_0 : \sigma^2 = \sigma_0^2$ versus $H_1 : \sigma^2 > \sigma_0^2$, $H_1 : \sigma^2 \neq \sigma_0^2$ ou $H_1 : \sigma^2 < \sigma_0^2$.
  Sous H_0, la statistique de test est :
  $$T_n = \frac{(n-1)S_n^2}{\sigma_0^2} \sim \chi^2_{n-1}$$
- *Conditions de validité.* Le caractère X est distribué suivant une loi normale.
- *Instruction R* : Il est possible d'utiliser la fonction `sigma2.test()` du *package* `sigma2tools`.
- Sans hypothèse de normalité et pour de grands échantillons, on peut utiliser la fonction `asymp.test(x,parameter= "var",reference=)`. Cette fonction se trouve dans le *package* `asympTest`.

Exemple 4.9.2. Considérons l'exemple d'une usine qui fabrique des boîtes de conserve de poids μ avec une précision $\sigma^2 = 10$. On veut montrer que la chaine de production est déréglée (précision différente de $\sigma^2 = 10$). Voici les poids d'une série de 20 boîtes de conserve :

165.1 171.5 168.1 165.6 166.8 170.0 168.8 171.1 168.8 173.6

163.5 169.9 165.4 174.4 171.8 166.0 174.6 174.5 166.4 173.8

```
poids <- c(165.1,171.5,168.1,165.6,166.8,170.0,168.8, 171.1,168.8,173.6,163.5,169.9,
165.4,174.4,171.8, 166.0,174.6,174.5,166.4,173.8)
sigma2.test(poids,var0=10)
```

4.9.5 Comparaison d'une proportion théorique à une valeur de référence

- *Descriptif du test.* Soit p la fréquence inconnue d'un caractère dans une population donnée. On observe des données de présence/absence de ce caractère sur les individus d'un échantillon de taille n de cette population. Les hypothèses du test que nous considérons sont : $H_0 : p = p_0$ versus $H_1 : p > p_0$, $H_1 : p \neq p_0$ ou $H_1 : p < p_0$.
  Sous H_0, la statistique de test est :
$$U = \frac{\widehat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim \mathcal{N}(0,1)$$
- *Conditions de validité.* L'échantillon doit être suffisamment grand, il faut vérifier que $np_0 \geq 5$ et $n(1-p_0) \geq 5$.
- *Instruction* **R** : Il est possible d'utiliser la fonction `prop.test()`.
- Sans hypothèse de normalité et pour de grands échantillons, on peut utiliser la fonction `asymp.test (x,parameter= "var",reference=)`. Cette fonction se trouve dans le *package* `asympTest`.
- *Cas des petits échantillons.* Dans ce cas, on peut effectuer un calcul exact fondé sur la loi binomiale grâce à la fonction `binom.test()`.

Exemple 4.9.3. Supposons que le recueil des données d'une étude de cas intitulée “Etude chez des femmes enceintes à Abidjan” ait été réalisé après une vaste campagne d'information et de prévention de l'infection par le VIH dont l'objectif était de réduire la proportion de personnes inféctées par le VIH, notamment chez les jeunes de 18 à 25 ans. Supposons encore que l'objectif à atteindre dans un premier temps soit la réduction de la prévalence dans cette population des femmes enceintes de 18 à 25 ans jusqu'à un taux inférieur à 10%. On veut donc savoir à l'aide du sous-échantillon des femmes de 25 ans ou moins si le taux de prévalence de l'infection par le VIH est inférieur à $p_0 = 0.1$. Le jeu de données se trouve à *http ://www.biostatisticien. eu/springeR/ prevalsidafric.xls.*

```
table(VIH[age<=25])
#  0   1
# 137 10
prop.test(10,147,0.1,alternative="less",correc=FALSE)
```

4.9.6 *Ce test n'a pas été traité en cours* : Comparaison d'un coefficient de corrélation théorique à une valeur de référence

- *Descriptif du test.* Soit ρ le coefficient de corrélation entre deux variables quantitatives X et Y. On souhaite tester les hypothèses : $H_0 \;:\; \rho = \rho_0$ versus $H_1 \;:\; \rho > \rho_0$, $H_1 \;:\; \rho \neq \rho_0$ ou $H_1 \;:\; \rho < \rho_0$.
  Sous H_0, la statistique de test est :
  $$U = \frac{Z - \mathbb{E}(Z)}{\sqrt{var(Z)}} \sim \mathcal{N}(0,1)$$
  avec
  $$Z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \;,\quad \mathbb{E}(Z) = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) \;,\quad var(Z) = \frac{1}{n-3}$$
  Dans le cas où l'on s'intéresse à l'association linéaire entre X et Y (testée en prenant $\rho_0 = 0$), la statistique de test utilisée, sous H_0, est :
  $$T + n = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}$$
- *Conditions de validité.* Le couple (X, Y) suit une loi binormale.
- *Instruction* **R** : Il est possible d'utiliser la fonction `cor.test()` dans le cas où l'on teste l'association linéaire entre X et Y. Sinon, pour une valeur autre que $\rho_0 = 0$, on peut utiliser la fonction `cor0.test()` disponible dans le *package* `LeLogicielR` associé à l'ouvrage de Lafaye de Micheaux et al. (2011).

Exemple 4.9.4. Pour la mise en oeuvre du test ci-dessus on peut considérer le jeu de données imc-enfant, où l'on s'intéresse à l'association linéaire entre la taille et le poids.

Exercice 42. Un producteur affirme qu'exactement 25% des haricots verts de sa récolte sont extra-fins. Sur 400 haricots verts choisis au hasard dans la récolte, on en compte 118 extra-fins.
Est-ce qu'on peut affirmer, au risque de 5%, que le producteur a tort ?

Exercice 43. Une entreprise utilise une matière isolante pour fabriquer des appareils de contrôle industriel. Elle achète des composants isolants à un certain fournisseur qui certifie que l'épaisseur moyenne de ses composants est de 7.3 millimètres. Pour voir si le fournisseur respecte ses engagements, l'entreprise mesure l'épaisseur de 24 composants pris au hasard dans la livraison. Les résultats, en millimètres, sont :

6.47 7.02 7.15 7.22 7.44 6.99 7.47 7.61 7.32 7.22 7.52 6.92

7.28 6.69 7.24 7.19 6.97 7.52 6.22 7.13 7.32 7.67 7.24 6.21

On suppose que l'épaisseur en millimètres d'un de ces composants peut être modélisée par une v.a.r. $X \sim \mathcal{N}(\mu, (0.38)^2)$, avec μ inconnu. Peut-on affirmer, avec un faible risque de se tromper, que le fournisseur ne respecte pas ses engagements ?

Exercice 44. Une usine fabrique un certain type de récipient en plastique. On cherche à montrer, avec un faible risque de se tromper, que le contenu moyen d'un récipient est strictement supérieur à 10 litres. Le contenu de 12 récipients choisis au hasard dans la production est mesuré. Les résultats, en litres, sont :

$$10.1 \quad 9.8 \quad 10.2 \quad 10.3 \quad 10.4 \quad 9.8 \quad 9.9 \quad 10.4 \quad 10.2 \quad 9.5 \quad 10.4 \quad 9.6$$

On suppose que le contenu en litres d'un récipient de cet usine peut être modélisé par une v.a.r. X suivant une loi normale. Proposer un test statistique adapté et conclure.

Exercice 45. Dans une production, pour que le poids annoncé du contenu d'une boîte de conserve de tomates soit conforme, il faut régler la moyenne du conditionnement à 276 grammes. Une panne est survenue dans la conditionneuse et le producteur craint que le réglage ne soit plus fiable. Il se pose la question : le réglage est-il encore à 276 grammes ?
Il prélève 8 boîtes au hasard dans la production et les pèse une à une. Les résultats, en grammes, sont :

$$232 \quad 277 \quad 235 \quad 245 \quad 245 \quad 250 \quad 268 \quad 256$$

On suppose que le poids en grammes du contenu d'une boîte de conserve de tomates de cette production peut être modélisé par une v.a.r. X suivant une loi normale. Faire un test statistique pour répondre à la question du producteur.

5.5 Travaux pratiques

La séance de TP se fait sous l'environnement Windows, sauf si l'on a une nette préférence pour Linux. Pour commencer la séance :

1. Créer un répertoire `TP_L3MIASHS_Stat2` sur le bureau.
2. Lancer ensuite `R` et modifier le répertoire de travail en allant dans `Fichier -> Changer le Répertoire Courant` et en choisissant le répertoire `Bureau/TP_L3MIASHS_Stat2` qui est déjà créé.
3. Ouvrir une fenêtre d'éditeur `Fichier -> Nouveau Script`.
4. Sauver le fichier dans le répertoire courant, par exemple, sous le nom `TP4.R : Fichier -> Sauver sous`
5. Pour les différentes questions, vous pouvez utiliser un "copier-coller" à partir de ce document. *Il est fortement recommandé de saisir toutes les commandes dans la fenêtre de l'éditeur que l'on a ouverte.* Pour exécuter les commandes saisies, il suffit de les selectionner avec la souris et d'appuyer simultanément sur les touches `Ctrl et R`.
6. Pour inclure des commentaires dans le programme, ce qui est fortement recommandé, on doit utiliser le caractère #. Tout ce qui suit le caractère # sera négligé lors de l'exécution.
7. Penser à sauvegarder régulièrement le contenu du fichier `TP4.R` en appuyant sur les touches `Ctrl et S`.

5.5.1 Tests d'homogénéité : échantillons indépendants

Nous rappelons, ci-dessous, quelques définitions et vocabulaires.

- *Contexte.* On étudie un caractère représenté par une v.a.r. X dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On cherche à comparer $\mathcal{P}_1$ et $\mathcal{P}_2$ quant à ce caractère, et donc à analyser leur éventuelle homogénéité. Pour ce faire, on considère :
  — un échantillon $\mathcal{E}_1$ de n_1 individus de $\mathcal{P}_1$,
  — un échantillon $\mathcal{E}_2$ de n_2 individus de $\mathcal{P}_2$.
- *Echantillons indépendants.* Si les individus sont différents, les échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$ sont indépendants.
- *Données.*
  — La v.a.r. X considérée dans $\mathcal{P}_1$ est une v.a.r. X_1,
  — La v.a.r. X considérée dans $\mathcal{P}_2$ est une v.a.r. X_2.
  Les données sont constituées de :
  — La valeur de X_1 pour chacun des n_1 individus de $\mathcal{E}_1$: $x_{1,1}, \ldots, x_{1,n}$,
  — La valeur de X_2 pour chacun des n_2 individus de $\mathcal{E}_2$: $x_{2,1}, \ldots, x_{2,n}$.
  On peut mettre les donnes sous la forme :
  — pour $\mathcal{E}_1$:
  $$x_{1,1} \quad x_{1,2} \quad \ldots \quad x_{1,n_1}$$
  — pour $\mathcal{E}_2$:
  $$x_{2,1} \quad x_{2,2} \quad \ldots \quad x_{2,n_2}$$

- *Formules : p-valeurs* (cf. Table 6.4).
  Lois :
  $$Z \sim \mathcal{N}(0,1),\ T \sim t_\nu \text{ et } K \sim \chi^2_\nu \text{ o } \nu = n-1$$
  Outils :
  $$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{et} \quad s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$
- *Commandes.* Pour les commandes (cf. Table 6.5), on considère les librairies `stats` et `OneTwoSamples` :
  ```
  library(stats)
  library(OneTwoSamples)
  ```

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$	H_1	Stat. test obs.	p-valeurs
σ_1, σ_2 connus : `2-Comp-Z-Test`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$z_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}}$	$\mathbb{P}(\|Z\| \geq \|z_{obs}\|)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
σ_1, σ_2 inconnus : `2-Comp-F-Test`	$\sigma_1^2 \neq \sigma_2^2$	$f_{obs} = \left(\dfrac{\max(s_1, s_2)}{\min(s_1, s_2)}\right)^2$	$2\,\mathbb{P}(F \geq f_{obs})$
σ_1, σ_2 inconnus, $\sigma_1 = \sigma_2$: `2-Comp-T-Test pooled yes`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\mathbb{P}(\|T_\nu\| \geq \|t_{obs}\|)$ $\mathbb{P}(T_\nu \geq t_{obs})$ $\mathbb{P}(T_\nu \leq t_{obs})$
σ_1, σ_2 inconnus, $\sigma_1 \neq \sigma_2$: `2-Comp-T-Test pooled no`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1}{n_1} + \dfrac{s_2}{n_2}}}$	$\mathbb{P}(\|T_\nu\| \geq \|t_{obs}\|)$ $\mathbb{P}(T_\nu \geq t_{obs})$ $\mathbb{P}(T_\nu \leq t_{obs})$
$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Stat. test obs.	p-valeurs
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$ $n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$: `2-Prop-Z-Test cor`	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z_{obs} = \dfrac{\|\bar{x}_1 - \bar{x}_2\| - \left(\dfrac{0.5}{n_1} + \dfrac{0.5}{n_2}\right)}{\sqrt{\bar{x}_p(1-\bar{x}_p)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$\mathbb{P}(\|Z\| \geq \|z_{obs}\|)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$ $n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$: `2-Prop-Z-Test`	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\bar{x}_p(1-\bar{x}_p)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$\mathbb{P}(\|Z\| \geq \|z_{obs}\|)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$

TABLE 5.4 – Tests de comparaison

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$	H_1	Commandes

σ_1, σ_2 connus :
`2-Comp-Z-Test`
	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`mean_test2(x1, x2, sigma = c(sigma1, sigma2))$p_value`
		`mean_test2(x1, x2, sigma = c(sigma1, sigma2), side = -1)$p_value`
		`mean_test2(x1, x2, sigma = c(sigma1, sigma2), side = 1)$p_value`

σ_1, σ_2 inconnus : `2-Comp-F-Test`	$\sigma_1^2 \neq \sigma_2^2$	`var.test(x1, x2)$p.value`

σ_1, σ_2 inconnus, $\sigma_1 = \sigma_2$:
`2-Comp-T-Test pooled yes`
	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`t.test(x1, x2, var.equal = T)$p.value`
		`t.test(x1, x2, alternative = "greater", var.equal = T)$p.value`
		`t.test(x1, x2, alternative = "less", var.equal = T)$p.value`

σ_1, σ_2 inconnus, $\sigma_1 \neq \sigma_2$:
`2-Comp-T-Test pooled no`
	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`t.test(x1, x2)$p.value`
		`t.test(x1, x2, alternative = "greater")$p.value`
		`t.test(x1, x2, alternative = "less")$p.value`

$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Commandes

$n_1 \geq 31$, $n_2 \geq 31$,
$n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$
$n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$:
`2-Prop-Z-Test cor`
	$p_1 \neq p_2$
	$p_1 > p_2$
	$p_1 < p_2$
		`prop.test(x = c(x1, x2), n = c(n1, n2))$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater")$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less")$p.value`

$n_1 \geq 31$, $n_2 \geq 31$,
$n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$
$n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$:
`2-Prop-Z-Test`
	$p_1 \neq p_2$
	$p_1 > p_2$
	$p_1 < p_2$
		`prop.test(x = c(x1, x2), n = c(n1, n2), correct = F)$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater", correct = F)$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less", correct = F)$p.value`

TABLE 5.5 – Commandes R pour les tests de conformité à une valeur de référence

Exercice 54. La société de Monsieur Labrador utilise deux machines : Machine 1 et Machine 2, pour remplir automatiquement des paquets de cacao en poudre.

— On prélève un échantillon de 10 paquets remplis par la Machine 1 et on les pèse. Les résultats, en grammes, sont :

106.70 107.02 107.15 107.22 107.41 106.39 107.47 107.61 107.38 107.22

— On prélève un échantillon de 9 paquets remplis par la Machine 2 et on les pèse. Les résultats, en grammes, sont :

107.68 106.69 107.24 107.69 106.97 107.52 106.22 107.23 107.32

On suppose que le poids, en grammes, d'un paquet rempli par la Machine 1 peut être modélisé par une v.a.r. $X_1 \sim \mathcal{N}(\mu_1, (1.3)^2)$ et celui avec la Machine 2 peut être modélisé par une v.a.r. $X_2 \sim \mathcal{N}(\mu_2, (0.9)^2)$. Peut-on affirmer, au risque de 5%, que les machines sont réglées de manière différente ?

Exercice 55. On dispose de deux lôts de boîtes de sauce italienne conditionnées de la même manière mais provenant de producteurs différents. On s'intéresse à la teneur, en grammes, en viande dans celles-ci.

— On extrait 7 boîtes provenant du premier producteur et on mesure leur teneur en viande. Les résultats, en grammes, sont :

2.12 12.03 13.58 13.38 11.81 15.92 13.65

— On extrait 6 boîtes provenant du deuxième producteur et on mesure leur teneur en viande. Les résultats, en grammes, sont :

14.81 13.93 14.91 15.87 15.62 15.39

La teneur, en grammes, en viande dans une boîte provenant du premier producteur peut être modélisée par une v.a.r. X_1, et celle dans une boîte provenant du deuxième producteur peut être modélisée par une v.a.r. X_2. On suppose que X_1 et X_2 suivent des lois normales.
Peut-on affirmer qu'il y a une différence entre les producteurs quant à la teneur moyenne en viande dans les boîtes ?

Exercice 56. On considère deux lôts de tasses et on souhaite comparer la solidité de celles-ci. Pour chacun des deux lots, on dispose d'un échantillon de 10 tasses et on mesure la résistance de chacune d'entre elles. Les résultats sont :

— pour le premier échantillon :

31.70 31.98 32.24 32.35 31.18 32.19 32.63 31.19 31.54 31.89

— pour le deuxième échantillon :

31.61 31.10 31.20 31.11 32.66 31.15 31.71 31.22 31.16 31.21

La solidité d'une tasse du premier lôt peut être modélisée par une v.a.r. X_1, et celle d'une tasse du second lôt peut être modélisée par une v.a.r. X_2. On suppose que X_1 et X_2 suivent des lois normales de variances égales. Peut-on affirmer que ces deux échantillons ne proviennent pas de la même production ?

Exercice 57. Un producteur de desserts lactés au caramel se trouve en concurrence avec d'autres marques. Au début de l'année 2010, il décide d'investir dans une nouvelle présentation de ses desserts. Avant d'avoir le bilan de l'année, il fait une rapide enquête auprès d'un certain nombre de magasins.

— Avant la nouvelle présentation, sur 230 desserts vendus, 54 étaient ceux du producteur.
— Après la nouvelle présentation, sur 340 desserts vendus, 110 étaient ceux du producteur.

Est-ce que le producteur peut affirmer que la nouvelle présentation a augmenté sa part de marché sur les desserts lactés au caramel ?

5.5.2 Tests d'homogénéité : échantillons appriés

- *Contexte.* On étudie un caractère, représenté par une v.a.r. X, dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On cherche à comparer $\mathcal{P}_1$ et $\mathcal{P}_2$ quant à ce caractère, et donc à analyser leur éventuelle homogénéité. Pour ce faire, on considère :
  — un échantillon $\mathcal{E}_1$ de n_1 individus de $\mathcal{P}_1$,
  — un échantillon $\mathcal{E}_2$ de n_2 individus de $\mathcal{P}_2$.
- *Echantillons appariés.* Si les individus de $\mathcal{P}_1$ sont soumis à un certain traitement (ou aucun), et ceux de $\mathcal{P}_2$ sont les individus de $\mathcal{P}_1$ soumis à un autre traitement, les échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$ sont appariés : ce sont les mêmes individus qui sont considérés dans les deux échantillons. On compare alors les effets des deux traitements en considérant un même échantillon de $n = n_1 = n_2$ individus.
- *Données.*
  — La v.a.r. X considérée dans $\mathcal{P}_1$ est une v.a.r. X_1,
  — La v.a.r. X considérée dans $\mathcal{P}_2$ est une v.a.r. X_2.
  Les données sont constituées de :
  — La valeur de X_1 pour chacun des $n_1 = n$ individus de $\mathcal{E}_1$: $x_{1,1}, \dots, x_{1,n}$,
  — La valeur de X_2 pour chacun des $n_2 = n$ individus de $\mathcal{E}_2$: $x_{2,1}, \dots, x_{2,n}$.
  Pour tout $i \in \{1, \dots, n\}$, sur le $i$ème individu, on observe donc une paire de valeurs : $(x_{1,i}, x_{2,i})$. Si on prend le schéma "Traitement 1" et "Traitement 2", on peut mettre les données sous la forme :
- *Formules : p-valeurs* (cf. Table 6.7).
  Lois dont on a besoin : $T \sim t_\nu$ où $\nu = n - 1$ et $K \sim \chi^2_1$.
  Outils :

$$\bar{d}_n = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad d_i = x_{1,i} - x_{2,i}, \quad s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n-1}(d_i - \bar{d})^2},$$

  pour tout $(i,j) \in \{0,1\}^2$, on pose $n_{i,j}$ le nombre d'individus dans l'échantillon vérifiant $X_1 = i$ et $X_2 = j$.

Individus	Traitement 1	Traitement 2
ω_1	$x_{1,1}$	$x_{2,1}$
ω_2	$x_{1,2}$	$x_{2,2}$
$\vdots$	$\vdots$	$\vdots$
ω_n	$x_{1,n}$	$x_{2,n}$

TABLE 5.6 – Schéma des données

— *Commandes.* On considère la librairie `stats` par la commande `library(stats)`. On propose donc les commandes `R` de la Table 6.8.

• *Cas de non normalité.* Pour la comparaison de 2 moyennes, si l'hypothèse que les v.a.r. suivent des lois normales ne semblent pas vérifiée (en faisant une analyse graphique ou un test statistique adéquat, comme le "test de Shapiro-Wilk"), certains tests dit "non paramétriques" proposent des alternatives satisfaisantes. Notamment, on peut utiliser :
— le test de Mann et Witney si les échantillons sont indépendants,
— le test de Wilcoxon si les échantillons sont appariés.
Les commandes associées sont :

```
wilcox.test(x,y)$p.value
wilcox.test(x,y,paired=TRUE)$p.value
```

Exercice 58. Un médecin ne veut se tromper que 5 fois sur 100 en décidant que l'administration d'un traitement particulier à un malade provoque, en moyenne, un accroissement de poids au bout de 3 mois de traitement. Le médecin examine le poids avant traitement et le poids après traitement de 5 malades choisis au hasard. Les résultats, en kilogrammes, sont :

Sujet n°	Poids avant traitement	Poids après traitement
1	80.82	83.76
2	60.12	64.13
3	102.52	101.81
4	51.65	56.63
5	65.96	68.21

Le poids en kilogrammes d'un malade avant traitement peut être modélisé par une v.a.r. X_1, et le poids en kilogrammes d'un malade après 3 mois de traitement peut être modélisé par une v.a.r. X_2. On suppose que $(X_1 - X_2)$ suit une loi normale. Proposer une modélisation du problème via un test statistique adapté et énoncer clairement la conclusion.

Exercice 59. La prise d'un médicament M_1, anti-inflammatoire, provoque quelquefois des douleurs gastriques. Le médecin propose la prise d'un médicament supplémentaire M_2 pour tenter d'éviter

$X_1 - X_2 \sim \mathcal{N}$ $\mathbb{E}(X_1) = \mu_1,\ \mathbb{E}(X_2) = \mu_2$	H_1	Stat. test obs.	p-valeurs
`Paired T-Test`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \sqrt{n}\left(\dfrac{\bar{d} - d_0}{s_n}\right)$	$\mathbb{P}\left(\lvert T\rvert \geq \lvert t_{obs}\rvert\right)$ $\mathbb{P}\left(T \geq t_{obs}\right)$ $\mathbb{P}\left(T \leq t_{obs}\right)$
$X_1 \sim \mathcal{B}(p_1),\ X_2 \sim \mathcal{B}(p_2)$	H_1	Stat. test obs.	p-valeurs
`MacNemar Test cor`	$p_1 \neq p_2$	$\chi^2_{obs} = \dfrac{\left(\lvert n_{0,1} - n_{1,0}\rvert - 1\right)^2}{n_{0,1} + n_{1,0}}$	$\mathbb{P}\left(K \geq \chi^2_{obs}\right)$
`MacNemar Test`	$p_1 \neq p_2$	$\chi^2_{obs} = \dfrac{\left(n_{0,1} - n_{1,0}\right)^2}{n_{0,1} + n_{1,0}}$	$\mathbb{P}\left(K \geq \chi^2_{obs}\right)$

TABLE 5.7 – Formules des p-valeurs

cet inconvénient. Ainsi, 87 malades présentant une infection inflammatoire et prenant le remède M_1 sont testés. On leur demande d'observer l'apparition ou non de douleurs gastriques avant et après l'administration du médicament supplémentaire M_2. Les résultats sont :

— 61 malades n'ont eu de douleurs gastriques ni avant ni après M_2

— 2 malades qui n'avaient pas eu de douleurs avant M_2 en ont eu après

— 11 malades qui ont eu des douleurs avant M_2 n'en ont plus eu après

— 13 malades ont eu des douleurs aussi bien avant qu'après M_2.

Peut-on affirmer que l'administration de M_2 a modifié la probabilité d'avoir des douleurs gastriques ?

$X - 1 - X_2 \sim \mathcal{N}$ $\mathbb{E}(X_1) = \mu_1,\ \mathbb{E}(X_2) = \mu_2$	H_1	Commandes

`Paired T-Test`	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`t.test(x1,x2,paired=T)$p.value`
		`t.test(x1,x2,paired=T,alternative="greater")$p.value`
		`t.test(x1,x2,paired=T,alternative="less")$p.value`

$X_1 \sim \mathcal{B}(p_1),\ X_2 \sim \mathcal{B}(p_2)$	H_1	Commandes
`MacNemar Test cor`	$p_1 \neq p_2$	`mcnemar(x1,x2)`
`MacNemar Test`	$p_1 \neq p_2$	`mcnemar(x1,x2,correct=F)`

TABLE 5.8 – Commandes `R` pour les tests de comparaison d'échantillons appariés

6.5 Travaux pratiques

La séance de TP se fait sous l'environnement Windows, sauf si l'on a une nette préférence pour Linux. Pour commencer la séance :

1. Créer un répertoire `TP_L3MIASHS_Stat2` sur le bureau.
2. Lancer ensuite `R` et modifier le répertoire de travail en allant dans `Fichier -> Changer le Répertoire Courant` et en choisissant le répertoire `Bureau/TP_L3MIASHS_Stat2` qui est déjà créé.
3. Ouvrir une fenêtre d'éditeur `Fichier -> Nouveau Script`.
4. Sauver le fichier dans le répertoire courant, par exemple, sous le nom `TP4.R : Fichier -> Sauver sous`
5. Pour les différentes questions, vous pouvez utiliser un "copier-coller" à partir de ce document. *Il est fortement recommandé de saisir toutes les commandes dans la fenêtre de l'éditeur que l'on a ouverte.* Pour exécuter les commandes saisies, il suffit de les selectionner avec la souris et d'appuyer simultanément sur les touches `Ctrl et R`.
6. Pour inclure des commentaires dans le programme, ce qui est fortement recommandé, on doit utiliser le caractère #. Tout ce qui suit le caractère # sera négligé lors de l'exécution.
7. Penser à sauvegarder régulièrement le contenu du fichier `TP4.R` en appuyant sur les touches `Ctrl et S`.

6.5.1 Tests d'homogénéité : échantillons indépendants

Nous rappelons, ci-dessous, quelques définitions et vocabulaires.

- *Contexte.* On étudie un caractère représenté par une v.a.r. X dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On cherche à comparer $\mathcal{P}_1$ et $\mathcal{P}_2$ quant à ce caractère, et donc à analyser leur éventuelle homogénéité. Pour ce faire, on considère :
  — un échantillon $\mathcal{E}_1$ de n_1 individus de $\mathcal{P}_1$,
  — un échantillon $\mathcal{E}_2$ de n_2 individus de $\mathcal{P}_2$.
- *Echantillons indépendants.* Si les individus sont différents, les échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$ sont indépendants.
- *Données.*
  — La v.a.r. X considérée dans $\mathcal{P}_1$ est une v.a.r. X_1,
  — La v.a.r. X considérée dans $\mathcal{P}_2$ est une v.a.r. X_2.
  Les données sont constituées de :
  — La valeur de X_1 pour chacun des n_1 individus de $\mathcal{E}_1$: $x_{1,1}, \ldots, x_{1,n}$,
  — La valeur de X_2 pour chacun des n_2 individus de $\mathcal{E}_2$: $x_{2,1}, \ldots, x_{2,n}$.
  On peut mettre les donnes sous la forme :
  — pour $\mathcal{E}_1$:
  $$x_{1,1} \quad x_{1,2} \quad \ldots \quad x_{1,n_1}$$
  — pour $\mathcal{E}_2$:
  $$x_{2,1} \quad x_{2,2} \quad \ldots \quad x_{2,n_2}$$

- *Formules : p-valeurs* (cf. Table 6.4).
  Lois :
  $$Z \sim \mathcal{N}(0,1),\ T \sim t_\nu \text{ et } K \sim \chi^2_\nu \text{ o } \nu = n-1$$
  Outils :
  $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- *Commandes.* Pour les commandes (cf. Table 6.5), on considère les librairies `stats` et `OneTwoSamples` :
  ```
  library(stats)
  library(OneTwoSamples)
  ```

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$	H_1	Stat. test obs.	p-valeurs
σ_1, σ_2 connus : `2-Comp-Z-Test`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$z_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}}}$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
σ_1, σ_2 inconnus : `2-Comp-F-Test`	$\sigma_1^2 \neq \sigma_2^2$	$f_{obs} = \left(\dfrac{\max(s_1, s_2)}{\min(s_1, s_2)}\right)^2$	$2\,\mathbb{P}(F \geq f_{obs})$
σ_1, σ_2 inconnus, $\sigma_1 = \sigma_2$: `2-Comp-T-Test pooled yes`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\mathbb{P}(\lvert T_\nu\rvert \geq \lvert t_{obs}\rvert)$ $\mathbb{P}(T_\nu \geq t_{obs})$ $\mathbb{P}(T_\nu \leq t_{obs})$
σ_1, σ_2 inconnus, $\sigma_1 \neq \sigma_2$: `2-Comp-T-Test pooled no`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1}{n_1} + \dfrac{s_2}{n_2}}}$	$\mathbb{P}(\lvert T_\nu\rvert \geq \lvert t_{obs}\rvert)$ $\mathbb{P}(T_\nu \geq t_{obs})$ $\mathbb{P}(T_\nu \leq t_{obs})$
$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Stat. test obs.	p-valeurs
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$ $n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$: `2-Prop-Z-Test cor`	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z_{obs} = \dfrac{\lvert\bar{x}_1 - \bar{x}_2\rvert - \left(\dfrac{0.5}{n_1} + \dfrac{0.5}{n_2}\right)}{\sqrt{\bar{x}_p(1-\bar{x}_p)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$
$n_1 \geq 31$, $n_2 \geq 31$, $n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$ $n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$: `2-Prop-Z-Test`	$p_1 \neq p_2$ $p_1 > p_2$ $p_1 < p_2$	$z_{obs} = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\bar{x}_p(1-\bar{x}_p)}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$\mathbb{P}(\lvert Z\rvert \geq \lvert z_{obs}\rvert)$ $\mathbb{P}(Z \geq z_{obs})$ $\mathbb{P}(Z \leq z_{obs})$

TABLE 6.4 – Tests de comparaison

$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$	H_1	Commandes

σ_1, σ_2 connus :
`2-Comp-Z-Test`
	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`mean_test2(x1, x2, sigma = c(sigma1, sigma2))$p_value`
		`mean_test2(x1, x2, sigma = c(sigma1, sigma2), side = -1)$p_value`
		`mean_test2(x1, x2, sigma = c(sigma1, sigma2), side = 1)$p_value`

σ_1, σ_2 inconnus : `2-Comp-F-Test`	$\sigma_1^2 \neq \sigma_2^2$	`var.test(x1, x2)$p.value`

σ_1, σ_2 inconnus, $\sigma_1 = \sigma_2$:
`2-Comp-T-Test pooled yes`
	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`t.test(x1, x2, var.equal = T)$p.value`
		`t.test(x1, x2, alternative = "greater", var.equal = T)$p.value`
		`t.test(x1, x2, alternative = "less", var.equal = T)$p.value`

σ_1, σ_2 inconnus, $\sigma_1 \neq \sigma_2$:
`2-Comp-T-Test pooled no`
	$\mu_1 \neq \mu_2$
	$\mu_1 > \mu_2$
	$\mu_1 < \mu_2$
		`t.test(x1, x2)$p.value`
		`t.test(x1, x2, alternative = "greater")$p.value`
		`t.test(x1, x2, alternative = "less")$p.value`

$X_1 \sim \mathcal{B}(p_1)$, $X_2 \sim \mathcal{B}(p_2)$	H_1	Commandes

$n_1 \geq 31$, $n_2 \geq 31$,
$n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$
$n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$:
`2-Prop-Z-Test cor`
	$p_1 \neq p_2$
	$p_1 > p_2$
	$p_1 < p_2$
		`prop.test(x = c(x1, x2), n = c(n1, n2))$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater")$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less")$p.value`

$n_1 \geq 31$, $n_2 \geq 31$,
$n_1\bar{x}_1 \geq 5$, $n_1(1-\bar{x}_1) \geq 5$
$n_2\bar{x}_2 \geq 5$, $n_2(1-\bar{x}_2) \geq 5$:
`2-Prop-Z-Test`
	$p_1 \neq p_2$
	$p_1 > p_2$
	$p_1 < p_2$
		`prop.test(x = c(x1, x2), n = c(n1, n2), correct = F)$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "greater", correct = F)$p.value`
		`prop.test(x = c(x1, x2), n = c(n1, n2), alternative = "less", correct = F)$p.value`

TABLE 6.5 – Commandes R pour les tests de conformité à une valeur de référence

Exercice 68. La société de Monsieur Labrador utilise deux machines : Machine 1 et Machine 2, pour remplir automatiquement des paquets de cacao en poudre.

— On prélève un échantillon de 10 paquets remplis par la Machine 1 et on les pèse. Les résultats, en grammes, sont :

106.70 107.02 107.15 107.22 107.41 106.39 107.47 107.61 107.38 107.22

— On prélève un échantillon de 9 paquets remplis par la Machine 2 et on les pèse. Les résultats, en grammes, sont :

107.68 106.69 107.24 107.69 106.97 107.52 106.22 107.23 107.32

On suppose que le poids, en grammes, d'un paquet rempli par la Machine 1 peut être modélisé par une v.a.r. $X_1 \sim \mathcal{N}(\mu_1, (1.3)^2)$ et celui avec la Machine 2 peut être modélisé par une v.a.r. $X_2 \sim \mathcal{N}(\mu_2, (0.9)^2)$. Peut-on affirmer, au risque de 5%, que les machines sont réglées de manière différente ?

Exercice 69. On dispose de deux lôts de boîtes de sauce italienne conditionnées de la même manière mais provenant de producteurs différents. On s'intéresse à la teneur, en grammes, en viande dans celles-ci.

— On extrait 7 boîtes provenant du premier producteur et on mesure leur teneur en viande. Les résultats, en grammes, sont :

2.12 12.03 13.58 13.38 11.81 15.92 13.65

— On extrait 6 boîtes provenant du deuxième producteur et on mesure leur teneur en viande. Les résultats, en grammes, sont :

14.81 13.93 14.91 15.87 15.62 15.39

La teneur, en grammes, en viande dans une boîte provenant du premier producteur peut être modélisée par une v.a.r. X_1, et celle dans une boîte provenant du deuxième producteur peut être modélisée par une v.a.r. X_2. On suppose que X_1 et X_2 suivent des lois normales.
Peut-on affirmer qu'il y a une différence entre les producteurs quant à la teneur moyenne en viande dans les boîtes ?

Exercice 70. On considère deux lôts de tasses et on souhaite comparer la solidité de celles-ci. Pour chacun des deux lots, on dispose d'un échantillon de 10 tasses et on mesure la résistance de chacune d'entre elles. Les résultats sont :

— pour le premier échantillon :

31.70 31.98 32.24 32.35 31.18 32.19 32.63 31.19 31.54 31.89

— pour le deuxième échantillon :

31.61 31.10 31.20 31.11 32.66 31.15 31.71 31.22 31.16 31.21

La solidité d'une tasse du premier lôt peut être modélisée par une v.a.r. X_1, et celle d'une tasse du second lôt peut être modélisée par une v.a.r. X_2. On suppose que X_1 et X_2 suivent des lois normales de variances égales. Peut-on affirmer que ces deux échantillons ne proviennent pas de la même production ?

Exercice 71. Un producteur de desserts lactés au caramel se trouve en concurrence avec d'autres marques. Au début de l'année 2010, il décide d'investir dans une nouvelle présentation de ses desserts. Avant d'avoir le bilan de l'année, il fait une rapide enquête auprès d'un certain nombre de magasins.

— Avant la nouvelle présentation, sur 230 desserts vendus, 54 étaient ceux du producteur.
— Après la nouvelle présentation, sur 340 desserts vendus, 110 étaient ceux du producteur.

Est-ce que le producteur peut affirmer que la nouvelle présentation a augmenté sa part de marché sur les desserts lactés au caramel ?

6.5.2 Tests d'homogénéité : échantillons appriés

- *Contexte.* On étudie un caractère, représenté par une v.a.r. X, dans deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$. On cherche à comparer $\mathcal{P}_1$ et $\mathcal{P}_2$ quant à ce caractère, et donc à analyser leur éventuelle homogénéité. Pour ce faire, on considère :
  — un échantillon $\mathcal{E}_1$ de n_1 individus de $\mathcal{P}_1$,
  — un échantillon $\mathcal{E}_2$ de n_2 individus de $\mathcal{P}_2$.
- *Echantillons appariés.* Si les individus de $\mathcal{P}_1$ sont soumis à un certain traitement (ou aucun), et ceux de $\mathcal{P}_2$ sont les individus de $\mathcal{P}_1$ soumis à un autre traitement, les échantillons $\mathcal{E}_1$ et $\mathcal{E}_2$ sont appariés : ce sont les mêmes individus qui sont considérés dans les deux échantillons. On compare alors les effets des deux traitements en considérant un même échantillon de $n = n_1 = n_2$ individus.
- *Données.*
  — La v.a.r. X considérée dans $\mathcal{P}_1$ est une v.a.r. X_1,
  — La v.a.r. X considérée dans $\mathcal{P}_2$ est une v.a.r. X_2.
  Les données sont constituées de :
  — La valeur de X_1 pour chacun des $n_1 = n$ individus de $\mathcal{E}_1$: $x_{1,1}, \dots, x_{1,n}$,
  — La valeur de X_2 pour chacun des $n_2 = n$ individus de $\mathcal{E}_2$: $x_{2,1}, \dots, x_{2,n}$.
  Pour tout $i \in \{1, \dots, n\}$, sur le $i$ème individu, on observe donc une paire de valeurs : $(x_{1,i}, x_{2,i})$. Si on prend le schéma "Traitement 1" et "Traitement 2", on peut mettre les données sous la forme :
- *Formules : p-valeurs* (cf. Table 6.7).
  Lois dont on a besoin : $T \sim t_\nu$ où $\nu = n - 1$ et $K \sim \chi^2_1$.
  Outils :

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i, \quad d_i = x_{1,i} - x_{2,i}, \quad s_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n-1}(d_i - \bar{d})^2},$$

  pour tout $(i, j) \in \{0, 1\}^2$, on pose $n_{i,j}$ le nombre d'individus dans l'échantillon vérifiant $X_1 = i$ et $X_2 = j$.

Individus	Traitement 1	Traitement 2
ω_1	$x_{1,1}$	$x_{2,1}$
ω_2	$x_{1,2}$	$x_{2,2}$
$\vdots$	$\vdots$	$\vdots$
ω_n	$x_{1,n}$	$x_{2,n}$

TABLE 6.6 – Schéma des données

— *Commandes.* On considère la librairie `stats` par la commande `library(stats)`. On propose donc les commandes `R` de la Table 6.8.

- *Cas de non normalité.* Pour la comparaison de 2 moyennes, si l'hypothèse que les v.a.r. suivent des lois normales ne semblent pas vérifiée (en faisant une analyse graphique ou un test statistique adéquat, comme le "test de Shapiro-Wilk"), certains tests dit "non paramétriques" proposent des alternatives satisfaisantes. Notamment, on peut utiliser :
  — le test de Mann et Witney si les échantillons sont indépendants,
  — le test de Wilcoxon si les échantillons sont appariés.
  Les commandes associées sont :
  ```
  wilcox.test(x,y)$p.value
  wilcox.test(x,y,paired=TRUE)$p.value
  ```

Exercice 72. Un médecin ne veut se tromper que 5 fois sur 100 en décidant que l'administration d'un traitement particulier à un malade provoque, en moyenne, un accroissement de poids au bout de 3 mois de traitement. Le médecin examine le poids avant traitement et le poids après traitement de 5 malades choisis au hasard. Les résultats, en kilogrammes, sont :

Sujet n°	Poids avant traitement	Poids après traitement
1	80.82	83.76
2	60.12	64.13
3	102.52	101.81
4	51.65	56.63
5	65.96	68.21

Le poids en kilogrammes d'un malade avant traitement peut être modélisé par une v.a.r. X_1, et le poids en kilogrammes d'un malade après 3 mois de traitement peut être modélisé par une v.a.r. X_2. On suppose que $(X_1 - X_2)$ suit une loi normale. Proposer une modélisation du problème via un test statistique adapté et énoncer clairement la conclusion.

Exercice 73. La prise d'un médicament M_1, anti-inflammatoire, provoque quelquefois des douleurs gastriques. Le médecin propose la prise d'un médicament supplémentaire M_2 pour tenter d'éviter

$X_1 - X_2 \sim \mathcal{N}$ $\mathbb{E}(X_1) = \mu_1,\ \mathbb{E}(X_2) = \mu_2$	H_1	Stat. test obs.	p-valeurs
`Paired T-Test`	$\mu_1 \neq \mu_2$ $\mu_1 > \mu_2$ $\mu_1 < \mu_2$	$t_{obs} = \sqrt{n}\left(\dfrac{\bar{d} - d_0}{s_n}\right)$	$\mathbb{P}\left(\lvert T\rvert \geq \lvert t_{obs}\rvert\right)$ $\mathbb{P}\left(T \geq t_{obs}\right)$ $\mathbb{P}\left(T \leq t_{obs}\right)$
$X_1 \sim \mathcal{B}(p_1),\ X_2 \sim \mathcal{B}(p_2)$	H_1	Stat. test obs.	p-valeurs
`MacNemar Test cor`	$p_1 \neq p_2$	$\chi^2_{obs} = \dfrac{\left(\lvert n_{0,1} - n_{1,0}\rvert - 1\right)^2}{n_{0,1} + n_{1,0}}$	$\mathbb{P}\left(K \geq \chi^2_{obs}\right)$
`MacNemar Test`	$p_1 \neq p_2$	$\chi^2_{obs} = \dfrac{\left(n_{0,1} - n_{1,0}\right)^2}{n_{0,1} + n_{1,0}}$	$\mathbb{P}\left(K \geq \chi^2_{obs}\right)$

TABLE 6.7 – Formules des p-valeurs

cet inconvénient. Ainsi, 87 malades présentant une infection inflammatoire et prenant le remède M_1 sont testés. On leur demande d'observer l'apparition ou non de douleurs gastriques avant et après l'administration du médicament supplémentaire M_2. Les résultats sont :

— 61 malades n'ont eu de douleurs gastriques ni avant ni après M_2
— 2 malades qui n'avaient pas eu de douleurs avant M_2 en ont eu après
— 11 malades qui ont eu des douleurs avant M_2 n'en ont plus eu après
— 13 malades ont eu des douleurs aussi bien avant qu'après M_2.

Peut-on affirmer que l'administration de M_2 a modifié la probabilité d'avoir des douleurs gastriques ?

$X-1-X_2 \sim \mathcal{N}$ $\mathbb{E}(X_1)=\mu_1,\ \mathbb{E}(X_2)=\mu_2$	H_1	Commandes

`Paired T-Test`	$\mu_1 \neq \mu_2$
	 $\mu_1 > \mu_2$
	 $\mu_1 < \mu_2$
		`t.test(x1,x2,paired=T)$p.value`
		 `t.test(x1,x2,paired=T,alternative="greater")$p.value`
		 `t.test(x1,x2,paired=T,alternative="less")$p.value`

$X_1 \sim \mathcal{B}(p_1),\ X_2 \sim \mathcal{B}(p_2)$	H_1	Commandes
`MacNemar Test cor`	$p_1 \neq p_2$	`mcnemar(x1,x2)`
`MacNemar Test`	$p_1 \neq p_2$	`mcnemar(x1,x2,correct=F)`

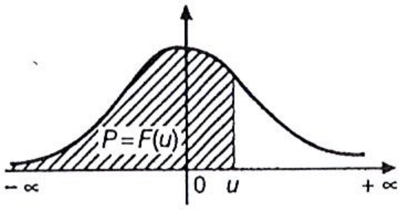
TABLE 6.8 – Commandes `R` pour les tests de comparaison d'échantillons appariés

Tables statistiques

Table 1

Fonction de répartition de la loi normale centrée réduite

Probabilité $F(u)$ d'une valeur inférieure à u



u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Tables pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
$F(u)$	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Table 2

Fractiles d'ordre P de la loi normale centrée réduite

P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,00	∞	3,0902	2,8782	2,7478	2,6521	2,5758	2,5121	2,4573	2,4089	2,3656	2,3263	0,99
0,01	2,3263	2,2904	2,2571	2,2262	2,1973	2,1701	2,1444	2,1201	2,0969	2,0749	2,0537	0,98
0,02	2,0537	2,0335	2,0141	1,9954	1,9774	1,9600	1,9431	1,9268	1,9110	1,8957	1,8808	0,97
0,03	1,8808	1,8663	1,8522	1,8384	1,8250	1,8119	1,7991	1,7866	1,7744	1,7624	1,7507	0,96
0,04	1,7507	1,7392	1,7279	1,7169	1,7060	1,6954	1,6849	1,6747	1,6646	1,6546	1,6449	0,95
0,05	1,6449	1,6352	1,6258	1,6164	1,6072	1,5982	1,5893	1,5805	1,5718	1,5632	1,5548	0,94
0,06	1,5548	1,5464	1,5382	1,5301	1,5220	1,5141	1,5063	1,4985	1,4909	1,4833	1,4758	0,93
0,07	1,4758	1,4684	1,4611	1,4538	1,4466	1,4395	1,4325	1,4255	1,4187	1,4118	1,4051	0,92
0,08	1,4051	1,3984	1,3917	1,3852	1,3787	1,3722	1,3658	1,3595	1,3532	1,3469	1,3408	0,91
0,09	1,3408	1,3346	1,3285	1,3225	1,3165	1,3106	1,3047	1,2988	1,2930	1,2873	1,2816	0,90
0,10	1,2816	1,2759	1,2702	1,2646	1,2591	1,2536	1,2481	1,2426	1,2372	1,2319	1,2265	0,89
0,11	1,2265	1,2212	1,2160	1,2107	1,2055	1,2004	1,1952	1,1901	1,1850	1,1800	1,1750	0,88
0,12	1,1750	1,1700	1,1650	1,1601	1,1552	1,1503	1,1455	1,1407	1,1359	1,1311	1,1264	0,87
0,13	1,1264	1,1217	1,1170	1,1123	1,1077	1,1031	1,0985	1,0939	1,0893	1,0848	1,0803	0,86
0,14	1,0803	1,0758	1,0714	1,0669	1,0625	1,0581	1,0537	1,0494	1,0450	1,0407	1,0364	0,85
0,15	1,0364	1,0322	1,0279	1,0237	1,0194	1,0152	1,0110	1,0069	1,0027	0,9986	0,9945	0,84
0,16	0,9945	0,9904	0,9863	0,9822	0,9782	0,9741	0,9701	0,9661	0,9621	0,9581	0,9542	0,83
0,17	0,9542	0,9502	0,9463	0,9424	0,9385	0,9346	0,9307	0,9269	0,9230	0,9192	0,9154	0,82
0,18	0,9154	0,9116	0,9078	0,9040	0,9002	0,8965	0,8927	0,8890	0,8853	0,8816	0,8779	0,81
0,19	0,8779	0,8742	0,8705	0,8669	0,8633	0,8596	0,8560	0,8524	0,8488	0,8452	0,8416	0,80
0,20	0,8416	0,8381	0,8345	0,8310	0,8274	0,8239	0,8204	0,8169	0,8134	0,8099	0,8064	0,79
0,21	0,8064	0,8030	0,7995	0,7961	0,7926	0,7892	0,7858	0,7824	0,7790	0,7756	0,7722	0,78
0,22	0,7722	0,7688	0,7655	0,7621	0,7588	0,7554	0,7521	0,7488	0,7454	0,7421	0,7388	0,77
0,23	0,7388	0,7356	0,7323	0,7290	0,7257	0,7225	0,7192	0,7160	0,7128	0,7095	0,7063	0,76
0,24	0,7063	0,7031	0,6999	0,6967	0,6935	0,6903	0,6871	0,6840	0,6808	0,6776	0,6745	0,75
0,25	0,6745	0,6713	0,6682	0,6651	0,6620	0,6588	0,6557	0,6526	0,6495	0,6464	0,6433	0,74
0,26	0,6433	0,6403	0,6372	0,6341	0,6311	0,6280	0,6250	0,6219	0,6189	0,6158	0,6128	0,73
0,27	0,6128	0,6098	0,6068	0,6038	0,6008	0,5978	0,5948	0,5918	0,5888	0,5858	0,5828	0,72
0,28	0,5828	0,5799	0,5769	0,5740	0,5710	0,5681	0,5651	0,5622	0,5592	0,5563	0,5534	0,71
0,29	0,5534	0,5505	0,5476	0,5446	0,5417	0,5388	0,5359	0,5330	0,5302	0,5273	0,5244	0,70
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Table 2 (suite)

Fractiles d'ordre P de la loi normale centrée réduite

P	0,000	0,001	0,002	0,003	0,004	0,005	0,006	0,007	0,008	0,009	0,010	
0,30	0,5244	0,5215	0,5187	0,5158	0,5129	0,5101	0,5072	0,5044	0,5015	0,4987	0,4959	0,69
0,31	0,4959	0,4930	0,4902	0,4874	0,4845	0,4817	0,4789	0,4761	0,4733	0,4705	0,4677	0,68
0,32	0,4677	0,4649	0,4621	0,4593	0,4565	0,4538	0,4510	0,4482	0,4454	0,4427	0,4399	0,67
0,33	0,4399	0,4372	0,4344	0,4316	0,4289	0,4261	0,4234	0,4207	0,4179	0,4152	0,4125	0,66
0,34	0,4125	0,4097	0,4070	0,4043	0,4016	0,3989	0,3961	0,3934	0,3907	0,3880	0,3853	0,65
0,35	0,3853	0,3826	0,3799	0,3772	0,3745	0,3719	0,3692	0,3665	0,3638	0,3611	0,3585	0,64
0,36	0,3585	0,3558	0,3531	0,3505	0,3478	0,3451	0,3425	0,3398	0,3372	0,3345	0,3319	0,63
0,37	0,3319	0,3292	0,3266	0,3239	0,3213	0,3186	0,3160	0,3134	0,3107	0,3081	0,3055	0,62
0,38	0,3055	0,3029	0,3002	0,2976	0,2950	0,2924	0,2898	0,2871	0,2845	0,2819	0,2793	0,61
0,39	0,2793	0,2767	0,2741	0,2715	0,2689	0,2663	0,2637	0,2611	0,2585	0,2559	0,2533	0,60
0,40	0,2533	0,2508	0,2482	0,2456	0,2430	0,2404	0,2378	0,2353	0,2327	0,2301	0,2275	0,59
0,41	0,2275	0,2250	0,2224	0,2198	0,2173	0,2147	0,2121	0,2096	0,2070	0,2045	0,2019	0,58
0,42	0,2019	0,1993	0,1968	0,1942	0,1917	0,1891	0,1866	0,1840	0,1815	0,1789	0,1764	0,57
0,43	0,1764	0,1738	0,1713	0,1687	0,1662	0,1637	0,1611	0,1586	0,1560	0,1535	0,1510	0,56
0,44	0,1510	0,1484	0,1459	0,1434	0,1408	0,1383	0,1358	0,1332	0,1307	0,1282	0,1257	0,55
0,45	0,1257	0,1231	0,1206	0,1181	0,1156	0,1130	0,1105	0,1080	0,1055	0,1030	0,1004	0,54
0,46	0,1004	0,0979	0,0954	0,0929	0,0904	0,0878	0,0853	0,0828	0,0803	0,0778	0,0753	0,53
0,47	0,0753	0,0728	0,0702	0,0677	0,0652	0,0627	0,0602	0,0577	0,0552	0,0527	0,0502	0,52
0,48	0,0502	0,0476	0,0451	0,0426	0,0401	0,0376	0,0351	0,0326	0,0301	0,0276	0,0251	0,51
0,49	0,0251	0,0226	0,0201	0,0175	0,0150	0,0125	0,0100	0,0075	0,0050	0,0025	0,0000	0,50
	0,010	0,009	0,008	0,007	0,006	0,005	0,004	0,003	0,002	0,001	0,000	P

Grandes valeurs de u

P	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}
u_P	3,7190	4,2649	4,7534	5,1993	5,6120	5,9978

Table 5

Fractiles d'ordre P de la loi χ^2_ν

ν \ P	0,005	0,01	0,025	0,05	0,1	0,25	0,5	0,7	0,75	0,9	0,95	0,975	0,99	0,995	0,999
1	0,000	0,0002	0,001	0,004	0,016	0,102	0,455	1,074	1,32	2,706	3,841	5,02	6,635	7,88	10,827
2	0,010	0,020	0,051	0,103	0,211	0,575	1,386	2,408	2,77	4,605	5,991	7,38	9,210	10,60	13,815
3	0,072	0,115	0,216	0,352	0,584	1,21	2,366	3,665	4,11	6,251	7,815	9,35	11,345	13,00	16,266
4	0,207	0,297	0,484	0,711	1,064	1,92	3,357	4,878	5,39	7,779	9,488	11,14	13,277	15,00	18,467
5	0,412	0,554	0,831	1,15	1,610	2,67	4,351	6,064	6,63	9,236	11,070	12,83	15,086	16,86	20,515
6	0,676	0,872	1,24	1,64	2,204	3,45	5,348	7,231	7,84	10,645	12,592	14,45	16,812	18,65	22,457
7	0,989	1,24	1,69	2,17	2,833	4,25	6,346	8,383	9,04	12,017	14,067	16,01	18,475	20,37	24,322
8	1,34	1,64	2,18	2,73	3,490	5,07	7,344	9,524	10,22	13,362	15,507	17,53	20,090	22,03	26,125
9	1,73	2,09	2,70	3,33	4,168	5,90	8,343	10,656	11,39	14,684	16,919	19,02	21,666	23,66	27,877
10	2,16	2,56	3,25	3,94	4,865	6,73	9,342	11,781	12,55	15,987	18,307	20,48	23,209	25,25	29,588
11	2,60	3,05	3,82	4,57	5,578	7,58	10,341	12,899	13,70	17,275	19,675	21,92	24,725	26,82	31,264
12	3,07	3,57	4,40	5,23	6,304	8,43	11,340	14,011	14,85	18,549	21,026	23,34	26,217	28,35	32,909
13	3,57	4,11	5,01	5,89	7,042	9,30	12,340	15,119	15,98	19,812	22,362	24,74	27,688	29,87	34,528
14	4,07	4,66	5,63	6,57	7,790	10,16	13,339	16,222	17,12	21,064	23,685	26,12	29,141	31,37	36,123
15	4,60	5,23	6,26	7,26	8,547	11,03	14,339	17,322	18,25	22,307	24,996	27,49	30,578	32,85	37,697
16	5,14	5,81	6,91	7,96	9,312	11,91	15,338	18,418	19,37	23,542	26,296	28,85	32,000	34,31	39,252
17	5,70	6,41	7,56	8,67	10,085	12,79	16,338	19,511	20,49	24,769	27,587	30,19	33,409	35,76	40,790
18	6,26	7,01	8,23	9,39	10,865	13,67	17,338	20,601	21,60	25,989	28,869	31,53	34,805	37,19	42,312
19	6,84	7,63	8,91	10,12	11,651	14,56	18,338	21,689	22,72	27,204	30,144	32,85	36,191	38,62	43,820
20	7,43	8,26	9,59	10,85	12,443	15,45	19,337	22,775	23,83	28,412	31,410	34,17	37,566	40,03	45,315
21	8,03	8,90	10,28	11,56	13,240	16,34	20,337	23,858	24,93	29,615	32,671	35,48	38,932	41,43	46,797
22	8,64	9,54	10,98	12,34	14,041	17,24	21,337	24,939	26,04	30,813	33,924	36,78	40,289	42,83	48,268
23	9,26	10,20	11,69	13,09	14,848	18,13	22,337	26,018	27,14	32,007	35,172	38,08	41,638	44,21	49,728
24	9,89	10,86	12,40	13,85	15,659	19,03	23,337	27,096	28,24	33,196	36,415	39,36	42,980	45,59	51,179
25	10,56	11,52	13,12	14,61	16,47	19,940	24,34	28,172	29,340	34,38	37,652	40,65	44,314	46,96	52,620
26	11,20	12,20	13,84	15,38	17,29	20,840	25,34	29,246	30,430	35,56	38,885	41,92	45,642	48,32	54,052
27	11,84	12,88	14,57	16,15	18,114	21,75	26,336	30,319	31,53	36,741	40,113	43,19	46,963	49,67	55,476
28	12,49	13,56	15,31	16,93	18,939	22,66	27,336	31,391	32,62	37,916	41,337	44,46	48,278	51,02	56,893
29	13,15	14,26	16,05	17,71	19,768	23,56	28,236	32,461	33,71	39,087	42,557	45,72	49,588	52,36	58,302
30	13,82	14,95	16,79	18,49	20,599	24,48	29,336	33,530	34,80	40,256	43,773	46,98	50,892	53,70	59,703
40	20,73	22,16	24,43	26,51	29,05	33,66	39,34	44,16	45,62	51,80	55,76	59,34	63,69	66,78	73,44
50	28,01	29,71	32,36	34,76	37,69	42,94	49,33	54,72	56,33	63,17	67,50	71,42	76,15	79,50	86,69
60	35,55	37,48	40,48	43,19	46,46	52,29	59,33	65,23	66,98	74,40	79,08	83,30	88,38	91,96	99,63
70	43,29	45,44	48,76	51,74	55,33	61,70	69,33	75,69	77,58	85,53	90,53	95,02	100,42	104,22	112,34
80	51,18	53,54	57,15	60,39	64,28	71,14	79,33	86,12	88,13	96,58	101,88	106,63	112,33	116,33	124,86
90	59,21	61,75	65,65	69,13	73,29	80,62	89,33	96,52	98,64	107,56	113,14	118,14	124,12	128,31	137,22
100	67,34	70,06	74,22	77,93	82,36	90,13	99,33	106,91	109,14	118,50	124,34	129,56	135,81	140,18	149,46

Table 6

Fractiles d'ordre P de la loi de Student T_ν

ν \ P	0,60	0,70	0,80	0,90	0,95	0,975	0,990	0,995	0,999	0,9995
1	0,325	0,727	1,376	3,078	6,314	12,71	31,82	63,66	318,3	636,6
2	0,289	0,617	1,061	1,886	2,920	4,303	6,965	9,925	22,33	31,60
3	0,277	0,584	0,978	1,638	2,353	3,182	4,541	5,841	10,22	12,94
4	0,271	0,569	0,941	1,533	2,132	2,776	3,747	4,604	7,173	8,610
5	0,267	0,559	0,920	1,476	2,015	2,571	3,365	4,032	5,893	6,859
6	0,265	0,553	0,906	1,440	1,943	2,447	3,143	3,707	5,208	5,959
7	0,263	0,549	0,896	1,415	1,895	2,365	2,998	3,499	4,785	5,405
8	0,262	0,546	0,889	1,397	1,860	2,306	2,896	3,355	4,501	5,041
9	0,261	0,543	0,883	1,383	1,833	2,262	2,821	3,250	4,297	4,781
10	0,260	0,542	0,879	1,372	1,812	2,228	2,764	3,169	4,144	4,587
11	0,260	0,540	0,876	1,363	1,796	2,201	2,718	3,106	4,025	4,437
12	0,259	0,539	0,873	1,356	1,782	2,179	2,681	3,055	3,930	4,318
13	0,259	0,538	0,870	1,350	1,771	2,160	2,650	3,012	3,852	4,221
14	0,258	0,537	0,868	1,345	1,761	2,145	2,624	2,977	3,787	4,140
15	0,258	0,536	0,866	1,341	1,753	2,131	2,602	2,947	3,733	4,073
16	0,258	0,535	0,865	1,337	1,746	2,120	2,583	2,921	3,686	4,015
17	0,257	0,534	0,863	1,333	1,740	2,110	2,567	2,898	3,646	3,965
18	0,257	0,534	0,862	1,330	1,734	2,101	2,552	2,878	3,611	3,922
19	0,257	0,533	0,861	1,328	1,729	2,093	2,539	2,861	3,579	3,883
20	0,257	0,533	0,860	1,325	1,725	2,086	2,528	2,845	3,552	3,850
21	0,257	0,532	0,859	1,323	1,721	2,080	2,518	2,831	3,527	3,819
22	0,256	0,532	0,858	1,321	1,717	2,074	2,508	2,819	3,505	3,792
23	0,256	0,532	0,858	1,319	1,714	2,069	2,500	2,807	3,485	3,767
24	0,256	0,531	0,857	1,318	1,711	2,064	2,492	2,797	3,467	3,745
25	0,256	0,531	0,856	1,316	1,708	2,060	2,485	2,787	3,450	3,725
26	0,256	0,531	0,856	1,315	1,706	2,056	2,479	2,779	3,435	3,707
27	0,256	0,531	0,855	1,314	1,703	2,052	2,473	2,771	3,421	3,690
28	0,256	0,530	0,855	1,313	1,701	2,048	2,467	2,763	3,408	3,674
29	0,256	0,530	0,854	1,311	1,699	2,045	2,462	2,756	3,396	3,659
30	0,256	0,530	0,854	1,310	1,697	2,042	2,457	2,750	3,385	3,646
32	0,256	0,530	0,853	1,309	1,694	2,037	2,449	2,738	3,365	3,622
34	0,255	0,529	0,852	1,307	1,691	2,032	2,441	2,728	3,348	3,601
36	0,255	0,529	0,852	1,306	1,688	2,028	2,434	2,719	3,333	3,582
38	0,255	0,529	0,851	1,304	1,686	2,024	2,429	2,712	3,319	3,566
40	0,255	0,529	0,851	1,303	1,684	2,021	2,423	2,704	3,307	3,551
50	0,255	0,528	0,849	1,298	1,676	2,009	2,403	2,678	3,261	3,496
60	0,254	0,527	0,848	1,296	1,671	2,000	2,390	2,660	3,232	3,460
70	0,254	0,527	0,847	1,294	1,667	1,994	2,381	2,648	3,211	3,435
80	0,254	0,527	0,846	1,292	1,664	1,990	2,374	2,639	3,195	3,415
90	0,254	0,526	0,846	1,291	1,662	1,987	2,368	2,632	3,183	3,402
100	0,254	0,526	0,845	1,290	1,660	1,984	2,365	2,626	3,174	3,389
200	0,254	0,525	0,843	1,286	1,653	1,972	2,345	2,601	3,131	3,339
500	0,253	0,525	0,842	1,283	1,648	1,965	2,334	2,586	3,106	3,310
∞	0,253	0,524	0,842	1,282	1,645	1,960	2,326	2,576	3,090	3,291

Table 7

Fractiles d'ordre 0,95 de la loi de Fisher-Snedecor $F(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	20	40	60	100	∞
1	161	200	216	225	230	234	237	239	242	244	248	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,79	8,74	8,66	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	5,96	5,91	5,80	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,74	4,68	4,56	4,46	4,43	4,41	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,06	4,00	3,87	3,77	3,74	3,71	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,64	3,57	3,44	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,35	3,28	3,15	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,14	3,07	2,94	2,83	2,79	2,76	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	2,98	2,91	2,77	2,66	2,62	2,59	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,85	2,79	2,65	2,53	2,49	2,46	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,75	2,69	2,54	2,43	2,38	2,35	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,67	2,60	2,46	2,34	2,30	2,26	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,60	2,53	2,39	2,27	2,22	2,19	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,54	2,48	2,33	2,20	2,16	2,12	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,49	2,42	2,28	2,15	2,11	2,07	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,45	2,38	2,23	2,10	2,06	2,02	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,41	2,34	2,19	2,06	2,02	1,98	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,38	2,31	2,16	2,03	1,98	1,94	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,35	2,28	2,12	1,99	1,95	1,91	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,32	2,25	2,10	1,96	1,92	1,88	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,30	2,23	2,07	1,94	1,89	1,85	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,27	2,20	2,05	1,91	1,86	1,82	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,25	2,18	2,03	1,89	1,84	1,80	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,24	2,16	2,01	1,87	1,82	1,78	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,22	2,15	1,99	1,85	1,80	1,76	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,20	2,13	1,97	1,84	1,79	1,74	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,19	2,12	1,96	1,82	1,77	1,73	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,18	2,10	1,94	1,81	1,75	1,71	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,16	2,09	1,93	1,79	1,74	1,70	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,08	2,00	1,84	1,69	1,64	1,59	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	1,99	1,92	1,75	1,59	1,53	1,48	1,39
120	3,92	3,07	2,68	2,44	2,29	2,17	2,08	2,01	1,91	1,83	1,65	1,49	1,42	1,36	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,83	1,75	1,57	1,39	1,32	1,24	1,00

Table 7 (suite)

Fractiles d'ordre 0,975 de la loi de Fisher-Snedecor $F(\nu_1, \nu_2)$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	10	12	20	40	60	100	∞
1	648	800	864	900	922	937	948	957	969	977	993	1006	1010	1013	1018
2	38,5	39,0	39,2	39,2	39,3	39,3	39,4	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,5
3	17,4	16,0	15,4	15,1	14,9	14,7	14,6	14,5	14,4	14,3	14,2	14,0	14,0	14,0	13,9
4	12,2	10,6	9,98	9,60	9,36	9,20	9,07	8,98	8,84	8,75	8,56	8,41	8,36	8,32	8,26
5	10,0	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,62	6,52	6,33	6,18	6,12	6,08	6,02
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,46	5,37	5,17	5,01	4,96	4,92	4,85
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,76	4,67	4,47	4,31	4,25	4,21	4,14
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,30	4,20	4,00	3,84	3,78	3,74	3,67
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	3,96	3,87	3,67	3,51	3,45	3,40	3,33
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,72	3,62	3,42	3,26	3,20	3,15	3,08
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,53	3,43	3,23	3,06	3,00	2,96	2,88
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,37	3,28	3,07	2,91	2,85	2,80	2,72
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,25	3,15	2,95	2,78	2,72	2,67	2,60
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,15	3,05	2,84	2,67	2,61	2,56	2,49
15	6,20	4,76	4,15	3,80	3,58	3,41	3,29	3,20	3,06	2,96	2,76	2,58	2,52	2,47	2,40
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	2,99	2,89	2,68	2,51	2,45	2,40	2,32
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,92	2,82	2,62	2,44	2,38	2,33	2,25
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,87	2,77	2,56	2,38	2,32	2,27	2,19
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,82	2,72	2,51	2,33	2,27	2,22	2,13
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,77	2,68	2,46	2,29	2,22	2,17	2,09
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,73	2,64	2,42	2,25	2,18	2,13	2,04
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,70	2,60	2,39	2,21	2,14	2,09	2,00
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,67	2,57	2,36	2,18	2,11	2,06	1,97
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,64	2,54	2,33	2,15	2,08	2,02	1,94
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,61	2,51	2,30	2,12	2,05	2,00	1,91
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,59	2,49	2,28	2,09	2,03	1,97	1,88
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,57	2,47	2,25	2,07	2,00	1,94	1,85
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,55	2,45	2,23	2,05	1,98	1,92	1,83
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,53	2,43	2,21	2,03	1,96	1,90	1,81
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,51	2,41	2,20	2,01	1,94	1,88	1,79
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,39	2,29	2,07	1,88	1,80	1,74	1,64
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,27	2,17	1,94	1,74	1,67	1,60	1,48
120	5,15	3,80	3,22	2,89	2,67	2,51	2,39	2,30	2,15	2,05	1,82	1,61	1,52	1,45	1,31
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,05	1,94	1,71	1,48	1,39	1,30	1,00