Quelques rudiments de Stat sous R.

(pour Statophobes)

Consigne : A la fin de la séance, envoyez moi un mail avec votre fichier R en pièce jointe, et avec comme sujet "TP R S3/ vos noms". Si vous écrivez du code dans la partie "console", il est exécuté à la volée mais sauvegardé nul part. Vous devez créer un script R, via le menu "File" puis "New File"puis "R Script" dans RStudio, que vous sauvegardez quelque part. Cliquez sur "Run" pour compiler la ligne où se trouve votre curseur.

1 Quelques rappels sur les lois sous R

1.1 Noms des fonctions

Si **xxx** désigne une loi de probablité alors **dxxx(), pxxx(), qxxx()** et **rxxx()** représentent respectivement la fonction de densité de probablité, la fonction de répartition, la réciproque de cette dernière et enfin la fonction de génération aléatoire de cette loi.

Par exemple, la loi normale est notée **norm**, nous avons donc :
- **dnorm()** avec d pour densité [1], qui représente la fonction de densité de probabilité de la loi normale.
- **pnorm()** avec p pour probabilité, qui représente la fonction de répartition de la loi normale.
- **qnorm()** avec q pour quantile, qui représente la fonction réciproque de la fonction de répartition de la loi normale.
- **rnorm()** avec r pour random (aléatoire), qui représente la fonction permettant de faire des tirages aléatoire selon une loi normale.

Cette convention de nommage est respectée par toutes les lois de probabilités définies dans R studio.

1.2 Nom des arguments : x, q, p n

Le premier argument des fonction **xxx** est toujours nommé de la façon suivante :
- **dxxx(x)** où x représente un vecteur de valeurs possibles pour une variable aléatoire suivant la loi **xxx**. Il peut être réduit à une seule valeur.
- **pxxx(q)** où q représente un (ou un vecteur) de valeur(s) de quantile(s).
- **qxxx(p)** où p représente une (ou un vecteur) de probabilités.
- **rxxx(n)** où n représente un entier, indiquant le nombre total de tirages aléatoire voulu

Avec ces notations on met bien en évidence les fonctions réciproques. On a par exemple :

```
qxxx(pxxx(q)) = q.
```

1.3 Quelques exemples

Taper ces exemples sous R studio pour comprendre le fonctionnement et vérifier au besoin avec un calcul ou les tables des lois (Normale, χ^2).
- La commande `dbinom(x=3, 10, 0.3)` renvoie $\mathbb{P}(Y = 3)$ où Y suit une Binomiale de paramètre $n = 10$ et $p = 0,3$. On peut aussi se contenter de taper `dbinom(3, 10, 0.3)`.
- `pbinom(3, 10, 0.3)` renvoie $\mathbb{P}(Y \leqslant 3)$ avec Y comme dans l'item précédent.
- `qbinom(0.6,10,0.3)` renvoie le plus petit entier k tel que $\mathbb{P}(Y \leqslant k) \geqslant 0,6$. C'est à dire 3 içi.
- Si l'on tape : `a=dpois(x=0:30,lambda=3)`, on obtient un vecteur à 31 coordonnées dont la ième contient la valeur $\mathbb{P}(X = i)$ où X est une v.a qui suit un loi de Poisson de paramètre $\lambda = 3$.

1. si la loi est discréte, cette commande renvoie la probabilité de tomber sur la valeur donnée

– `runif(10,0,20)` renvoie 10 valeurs tirées selon la loi uniforme sur l'intervalle $[0; 20]$.

– `pnorm(3)` renvoie $\mathbb{P}(Z \leqslant 3)$ où Z suit une $\mathcal{N}(0; 1)$. On peut également taper `pnorm(3,mean=0,sd=1)` ou `pnorm(3,0,1)` ou encore `pnorm(q=3,mean=0,sd=1)`

– Pour une loi normale générale, la commande `pnorm(q=9,mean=5,sd=2)` renvoie $\mathbb{P}(X \leqslant 9)$ où X suit une $\mathcal{N}(5; 2)$. Là encore, on peut se contenter de taper `pnorm(9,5,2)`.

– La commande `>plot(dnorm(x=20:100,mean=60,sd=10))` vous renvoie l'allure discrétisée de la densité d'une $N(60; 10)$.

– La commande `>qchisq(0.9, 5)` vous renvoie le nombre z (quantile d'un χ^2 à 5 ddl) tel que $\mathbb{P}(\chi < z) = 0.9$ où $\chi \sim \chi^2(5)$.

Exercice 1

1. Créez 500 notes distribuées suivant une loi Normale de moyenne de 10 et d'écart type de 2. A l'aide des fonctions `plot()` et `hist`, vérifiez graphiquement que les notes suivent bien la loi Normale.
2. Soit $Y \sim Binomiale(40, 0.5)$. Calculer $\mathbb{P}(Y = 15)$, puis $\mathbb{P}(Y \leqslant 15)$. Comparer avec $\mathbb{P}(Z \leqslant 15)$ où $Z \sim \mathcal{N}(20, \sqrt{10})$.

2 Régression linéaire (par les moindres carrés)

On souhaite savoir si il y a une dépendance (linéaire) entre deux séries $x = (x_i)_{i=1..n}$ et $y = (y_i)_{i=1..n}$, et si oui, connaître cette dépendance. Voilà quelques étapes pour répondre à ce genre de questions sous R.

– Etape 1 : Saisir les séries. Assigner par exemple :
`>x=c(0,1,2,5,7); y=c(1,4,7,16,22)`
Nous avons choisi un ex où la dépendance est "claire", puisque l'on a $y = 3x + 1$. Essayons de retrouver le 3 et le 1.

– Etape 2 : Représentation graphique. Cela permet de se donner une idée à moindre coût de l'alignement potentiel des points du nuage. On utilise la fonction plot() :

`>plot(x, y)` ou `>plot(x, y,pch=3)` (si l'on préfére que les points soient représentés par des croix plutôt que des ronds)

– Etape 3 : Calcul du coefficient de correlation. Il suffit de taper : `>cor(x,y)`

– Etape 4 : Equation de la droite et tracé. Sous R, la fonction `lm()` (pour linear model) nous permet d'obtenir l'équation de la droite d'ajustement.
`>lm(y~ x )`
Attention à l'ordre des séries dans cette commande. *Intercept* correspond à l'ordonnée à l'origine, et le x correspond au coefficient directeur. On peut tracer cette droite par la commande :
`>abline(lm(y~ x), col = 'red')`
Le 1er argument de abline correspond à l'ordonnée à l'origine, et le 2nd au coefficient directeur.

On peut alors utiliser le modèle pour faire des prédictions avec la commande `predict()` plutôt que de taper le calcul à la main... Supposons par exemple, que l'on cherche la valeur de la variable y lorsque $x = 6$. On peut taper sous R :

```
>y_pred <- predict( lm(y~x), newdata = data.frame(x =6))
```

Puis, on demande à R la valeur :

```
> y_pred
1
19
```

Exercice 2

Une entreprise livre des produits conditionnés en colis cartonnés. On a observé l'évolution du nombre de colis livrés par l'entreprise entre 1989 et 1996 :

Année	Rang de l'année x_i	Nombre de colis q_i
1997	1	7332
1998	2	8249
1999	3	8838
2000	4	9280
2001	5	9639
2002	6	9943
2003	7	10187
2004	8	10402

1. Calculer le coefficient de corrélation entre x_i et q_i, et l'équation de la droite d'ajustement associée. Représenter le nuage de points et la droite obtenue.
2. On pose $y_i = \ln(x_i)$ et $z_i = \ln(q_i)$, où ln désigne le logarithme népérien. Déterminer les valeurs de y_i et z_i. Représenter le nuage de points $M(y_i; z_i)$ et la courbe d'ajustement qui semble la mieux adaptée.
3. Calculer les coefficients de corrélation linéaire entre y_i et z_i . Que peut on déduire des valeurs obtenues ?
4. Grâce à vos conclusions, déterminer une expression de q en fonction de x (la plus adaptée) . Estimer alors le nombre de colis qui seront livrés par cette entreprise en 1997, en 2050. Commenter ces résultats

Exercice 3

Dans le tableau suivant, on donne la population du Canada de 1950 à 1995. La population est donnée en millions d'habitants.

Années	1950	1960	1970	1980	1990	1995
Population au Canada	$13,1$	$17,7$	21	$23,8$	$26,2$	$27,6$

Trois modèles d'ajustement sont envisagés pour cette série :
- Linéaire : $y = ax + b$
- Puissance : $y = a\sqrt{x} + b$
- Logarithmique : $y = a\ln(x) + b$.

1. A l'aide de changements de variables permettant un ajustement linéaire, déterminer les coefficients de correlation dans chacun des cas.
2. Tracer les 3 nuages de points (associés aux variables trouvées précédemment) sur 3 graphiques différents. Tracer les droites d'ajustements associées
3. Quel est le meilleur modèle selon vous ? Tracer sur un nouveau graphique le nuage de points "brut", ainsi que que la fonction qui vous semble approcher le mieux ces données.
4. En déduire une prévision de la population au Canada en 2020.

Exercice 4

Générer une suite de 50 nombres aléatoires compris entre 0 et 999. Soit x_i le chiffre des centaines et y_i le nombre formé par les deux derniers chiffres (par exemple pour 458, on a $x_1 = 4$ et $y_1 = 58$). Présenter vos résultats dans un tableau. (On pourra utiliser la fonction `floor(x)` de R qui renvoie la partie entière de x.

1. Déterminer l'équation de la droite d'ajustement de $Y = (y_i)_i$ en $X = (x_i)_i$ par la méthode des moindres carrés.
2. Calculer le coefficient de corrélation linéaire et interpréter le résultat.
3. Représenter graphiquement le nuage et la droite d'ajustement

3 Tests Statistiques

3.1 Tests de conformité de moyenne et intervalles de confiance

Pour ce qui suit, charger les librairies stats et OneTwoSamples :

```
>library(stats)
>library(OneTwoSamples)
>library(TeachingDemos)
```

On suppose donné une série d'observations $(X_i)_{i=1..n}$, iid d'espérance μ inconnue et d'écart-type σ (connu ou inconnu). Soit μ_0 une valeur présentie (ou pas) pour être l'espérance μ. On veut tester l'hyp H_0 = "$\mu = \mu_0$" contre une hyp H_1 à définir.

Si σ est connu	Si σ est inconnu
Nom du test	
Test Z de l'écart-réduit	Test T de Student
Commande	
`>z.test()` [1]	`>t.test()`
Conditions	
$n \geqslant 30$	$n \leqslant 30$ et les X_i Gaussiens
Syntaxe et Arguments	
`>z.test(x,mu=`μ_0`,sd=`σ`, alternative="...",conf.level=...)`	`>t.test(x, mu =`μ_0`,alternative="...",conf.level=...)`
`x` représente le vecteur "série des observations"	
`alternative` réprésente H_1, qui peut être : "`less`" pour $H_1 =$ "$\mu < \mu_0$. Test unilatéral "`two.sided`" pour $H_1 =$ "$\mu \neq \mu_0$". Test bilatéral "`greater`" pour$H_1 =$ "$\mu > \mu_0$. Test unilatéral	
`conf.level` représente le niveau de confiance du test, par ex 0.95	

Le résultat de ces deux commandes, est une liste contenant, entre autres, les éléments suivants :
- la valeur de la statistique `t`,
- le degré de liberté `df`,
- la `p-value` du test (la probabilité sous H_0, d'un résultat au moins aussi extrême que le résultat observé) [2]
- `conf.int` réprésente l'intervalle de confiance de la moyenne (à 95% par défaut).

Un exemple éclaircira les choses.

Exemple 1 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 $\mathcal{N}(\mu; 0.38)$ avec μ inconnu. Peut-on affirmer, avec un faible risque de se tromper, que le fournisseur ne respecte pas ses engagements ?

Eléments de réponse On connaît l'écart-type et on dispose de $n = 24$ observations. $n \leqslant 30$ mais les v.a sont Gaussiennes donc on peut faire le test Z bilatéral (à 5%). On pose

$$H_0 = "\mu = 7.3" \text{ contre } H_1 = "\mu \neq 7.3"$$

On rentre d'abord sous R les observations dans une serie nommée x :

```
> x=c(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)
```

Puis,

```
> z.test(x,mu=7.3, sd=0.38, alternative="two.sided", conf.level=0.95)
```

L'argument `conf.level` est facultatif. En effet, R prend par défaut un seuil de 5%. R renvoie :

```
One Sample z-test

data: x
z = -2.24, n = 24.000000, Std. Dev. = 0.380000, Std. Dev. of the sample mean = 0.077567,
p-value = 0.02509
```

1. La commande `mean_test1()` fonctionne également.
2. Ce nombre est utilisé pour conclure sur le résultat d'un test statistique. La procédure généralement employée consiste à comparer la *p-value* à un seuil préalablement défini (traditionnellement 5%). Si la *p-value* est inférieure à ce seuil, on rejette l'hypothèse nulle en faveur de l'hypothèse alternative, et le résultat du test est déclaré statistiquement significatif . Dans le cas contraire, si la *p-value* est supérieure au seuil, on ne rejette pas l'hypothèse nulle.

```
alternative hypothesis: true mean is not equal to 7.3
95 percent confidence interval:
6.974221 7.278279
sample estimates:
mean of x 7.12625
```

La valeur de la statistique $U = \frac{\sqrt{n}}{\sigma}(\overline{X_n} - 7.3)$ vaut -2.24 et la `p-value` vaut $0.025 < 0.05$, donc on rejette H_0 et le fournisseur ne respecte pas ses engagements.

Exemple 2 Dans une usine, une machine automatisée remplit des récipients en plastique. On cherche à montrer, avec un faible risque de se tromper, que le contenu moyen injectée par la machine dans le 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 :

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

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

Eléments de réponse

Sans ambiguïté, ne connaîssant pas l'écart-type, on utilise un test T de Student unilatéral (à 5%). Notons X_i le volume mesuré du récipient i, d'espérance μ inconnue. On pose

$$H_0 = "\mu = 10" \text{ contre } H_1 = "\mu > 10"$$

Dans R, nous rentrons les données puis la commande `t.test` :

```
>x = c(10.1, 9.8, 10.2, 10.3, 10.4, 9.8, 9.9, 10.4, 10.2, 9.5, 10.4, 9.6)
>t.test(x, mu = 10, alternative = "greater")
```

Cela renvoie :

```
One Sample t-test

data: x
t = 0.5404, df = 11, p-value = 0.2998
alternative hypothesis: true mean is greater than 10
95 percent confidence interval:
9.883838 Inf
sample estimates:
mean of x
10.05
```

La statistique $T = \frac{\sqrt{n-1}}{S_n}(\overline{X_n} - \mu_0)$ vaut ici 0.5404. Le paramètre de la loi (limite) de Student est 11, et la `p-value` vaut $0.2998 > 0.05$. Donc, on ne rejette pas H_0, les données ne nous permettent pas d'affirmer que le contenu moyen des récipients de cette usine est strictement supérieur à 10 litres.

Exercice 5 *La vie de Mr Slow...*

Dans une entreprise de conditionnement de colis, chaque employé est supposé, s'occuper de 45 colis par jours. Le chef de service soupçonne un employé *Mr Slow*, de travailler lentement et il effectue quelques mesures à son insu. Il note, sur une période de 15 jours, le nombre de colis qu'il traite quotidiennement. Il obtient les résultats suivants :

$$44, 38, 45, 46, 34, 39, 43, 40, 44, 48, 46, 41, 43, 44, 39.$$

Peut on considèrer que *Mr Slow*, est plus lent que ses collègues de travail (au risque 5%).
(*On supposera que le nombre de colis traités par un employé suit une loi normale*)

Exercice 6

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 :

$$232, 277, 235, 245, 245, 250, 268, 256$$

On suppose que le poids en grammes du contenu d'une boîte de conserve de tomates de cette production suit une loi Normale. Faire un test statistique pour répondre à la question du producteur.

Exercice 7

Une entreprise de camions dispose de 100 véhicules. Sur un échantillon de 31 jours, elle note le nombre de camions en panne

$$5,\ 5\ ,\ 6,\ 4,\ 6,\ 6,\ 8,\ 3,\ 5,\ 5,\ 5,\ 4,\ 3,\ 6,\ 5,\ 6,\ 4, 7,\ 6,\ 6,\ 5,\ 4,\ 3,\ 6,\ 5,\ 4,\ 5,\ 4,\ 5,\ 5,\ 1.$$

1. Calculer la moyenne et l'écart-type de cet ?échantillon.
2. En déduire une estimation ponctuelle de la moyenne et de l'écart-type du nombre de pannes sur l'ensemble des journées de l'année.
3. Déterminer un intervalle de confiance pour cette moyenne avec un coefficient de confiance de 95%.

Indication pour les exercices suivants. Lorsqu'on a accès directement à la moyenne empirique $\overline{X_n}$ sans la série, il suffit de remplacer x par la valeur de cette moyenne, de spécifier l'écart-type par $sd = ...$ et la taille de l'échantillon par $n = ...$

Exercice 8 *Proportions de cOoLs*

On suppose que moins de 20% de tous les travailleurs sont prêts à moins travailler et à être moins payés pour avoir plus de loisirs personnels. Un sondage aux USA révéle que sur un échantillon de taille 596, 83 personnes étaient prêtes à travailler moins pour un salaire moins important afin d'avoir plus de loisirs personnels. Notons p la "vraie" proportion de travailleurs prêts à moins travailler et à être moins payés pour avoir plus de loisirs personnels. Testez l'hypothèse H_0 : "$p = 20\%$" contre H_1 : "$p < 20\%$" au niveau $\alpha = 0,05$.

Exercice 9 *Haricots verts extra fins*

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 que l'on peut affirmer, au risque 5%, que le producteur a tort ?

3.2 Intervalles de confiance

Les affichages des commandes précédentes contiennent les IC. Pour y avoir accès directement, il suffit d'ajouter à la fin de la ligne de commande `$conf.int` Revenons à l'exemple 1. Si l'on cherche l'intervalle de confiance au taux de 0.95, de l'épaisseur des composants, on tapera :

```
>z.test(x,mu=7.3, sd=0.38, alternative="two.sided", conf.level=0.95)$conf.int
```

Remarque : Pour l'obtention d'intervalles de confiance de proportions, il existe une multitude de commandes spécifiques dues à des variantes de technologies mathématiques (par ex : méthode des scores de Wilson avec correction de continuité) comme `prop.test()` ou même des méthodes en utilisant des lois exactes (binomiale) comme `binom.test()`

Pour l'exercice 7, la commande :

```
> prop.test(83,596,conf.level=0.95)$conf.int
```

donnera un IC légèrement différent que la commande :

```
>z.test(83/596, mu = 0.2, sd=0.5, alternative = "two.sided",n=596)$conf.int.
```

On a mis `sd=0.5`, puisque on majore classiquement l'écart-type $\sqrt{\mu(1-\mu)}$ par $1/2$. Vérifiez que l'intervalle de confiance de cette deuxième commande correspond bien à celui du cours en tapant :

```
>83/596 -qnorm(0.975) *1/(2*sqrt(596))
```

et `>83/596 +qnorm(0.975) *1/(2*sqrt(596))`

3.3 Tests de conformité de variance

En spécifiant les instructions à la main ! Traitons par exemple le cas d'un test de variance bilatéral, avec une espérance inconnue. Soit x une série dont la variance présentie est σ_0^2. On veut tester l'hyp H_0 = "$\sigma = \sigma_0$" contre l'hyp H_1 = "$\sigma \neq \sigma_0$" au risque α. Prenons des valeurs numériques et entrons les dans R :

```
>x=c(10,14,6,5,15,7)
```

```
>alpha=0.05
>sigmao=4  (c'est l'hyp H_0)
```

On calcule alors la statistique du test $n\frac{S_n^2}{\sigma_0^2}$ et on la compare aux quantiles du $\chi^2(n)$, où `n=length(x)` vaut 6 ici. (Attention, les fonctions `var` et `sd()` de R, sont les estimateurs sans biais.) On tape donc :

```
>z1=qchisq(1-alpha/2,length(x)); z2=qchisq(alpha/2,length(x))
>Z= (length(x)-1)* sd(x)^2/sigmao^2

>if (Z < z1 & Z>z2) print("on  ne  rejette  pas  Ho") else print("on  rejette  Ho")
```

On peut condenser sans variables intermédiaires en :

```
>if ((length(x)-1)*sd(x)^ 2/sigmao ^ 2 <qchisq(1-alpha/2,length(x)) & (length(x)-1)*sd(x)
^ 2/sigmao ^ 2 > qchisq(alpha/2,length(x))) print("on  ne  rejette  pas  Ho") else
print("on  rejette  Ho")
```

3.4 Tests du χ^2

On utilise la fonction `>chisq.test()`. Nous allons voir directement sur des exemples comment fonctionnent ces commandes. A noter, que le niveau de confiance ne se spécifie pas dans les arguments de la fonction. On se le fixe au départ, et on le compare à la `p-value` que renvoie le test.

3.4.1 Tests du χ^2 d'adéquation

Exemple 3 La charte d'une grosse entreprise est de fonctionner de façon optimale avec les proportions d'employés suivantes : 15% de cadres, 30% de commerciaux et le reste d'ouvriers. On observe en réalité sur un échantillon de 54 employés : 8 cadres, 17 commerciaux et 29 ouvriers. Est ce que cet échantillon est conforme à la charte de l'entreprise (au taux de 0.95) ?

Eléments de réponse

On pose comme hypothèse nulle H_0 ="l'échantillon est conforme à la charte". On effectue donc un test du χ^2 d'adéquation à la loi discréte suivante :

statut	cadres	commerciaux	ouvrier
proba	0.15	0.3	0.55

On entre dans R les valeurs observées ainsi que les probabilités théoriques :

```
>Vobs=c(8,17,29)
>Ptheo=c(0.15,0.3,0.55)
```

Pour réaliser le test, on tape alors :

```
>chisq.test(Vobs,p=Ptheo)
```

On obtient :

```
Chi-squared test for given probabilities

data: Vobs
X-squared = 0.057239, df = 2, p-value = 0.9718
```

On peut vérifier "à la main" la valeur de la statistique donné sous le nom `X-squared` :

```
>Ntheo=c(0.15*54,0.3*54,0.55*54)
> Ntheo
[1] 8.1 16.2 29.7
> U=(8.1-8)^2/8.1+(16.2-17)^2/16.2+(29.7-29)^2/29.7
> U
[1] 0.05723906
```

3.4.2 Tests du χ^2 d'indépendance et d'homogénité

C'est une mini variante de la fonction précédente. Nous la traiterons par un exemple.

Exemple 4 Intéressons-nous aux salaires des hommes et des femmes. Une étude portant sur 290 hommes et 285 femmes, a fourni les résultats suivants :

Genre \ Salaire (en €)	1000 à 2000	2000 à 3000	3000 à 4000	>4000
Hommes	50	70	110	60
Femmes	80	75	100	30

Vérifier si les hommes et les femmes ont effectivement le même salaire.

On effectue donc un test du χ^2 d'homogénéité avec comem hyp nulle $H_0 =$ "les traitements H, F sont identiques" On entre d'abord les valeurs dans R. Pour cela on fabrique une matrice correspondante au tableau de l'énoncé :

```
>données=matrix(c(50 ,70 , 110 , 60, 80 , 75 , 100 , 30), nrow = 2, byrow = TRUE)
```

R calcule la statistique du χ^2 par la même commande que la section précédente :

```
>chisq.test(données)

Pearson's Chi-squared test
data: données
X-squared = 17.53, df = 3, p-value = 0.0005499
```

On peut obtenir le tableau des effectifs théoriques, le tableau des effectifs observés (celui de l'énoncé) et enfin le tableau des différences ($\frac{eff_{obs}-eff_{theo}}{\sqrt{eff_{theo}}}$) par les fonctions suivantes :

```
>chisq.test(données)$expected
        [,1]     [,2]    [,3]    [,4]
[1,] 65.56522 73.13043 105.913 45.3913
[2,] 64.43478 71.86957 104.087 44.6087
> chisq.test(données)$observe
     [,1] [,2] [,3] [,4]
[1,]   50   70  110   60
[2,]   80   75  100   30
> chisq.test(données)$residual
          [,1]       [,2]       [,3]      [,4]
[1,] -1.922288 -0.3660628  0.3971232  2.168329
[2,]  1.939077  0.3692599 -0.4005916 -2.187266
```

Exercice 10

Le tableau ci-dessous donne le nombre de cv qui ont retenu l'attention et ceux qui ont été écartés dans 3 entreprises A,B et C, pour un poste similaire.

Statut du cv \ Entreprise	A	B	C
cv écartés	50	47	56
cv retenus	5	14	8

Peut-on affirmer, au risque 5%, que le résultat d'un cv dépend de l'entreprise ?

Exercice 11

Un commercial fournissant les stations-service souhaite savoir s'il y a un lien entre l'achat de bière et l'achat de paquets de chips. Pour le tester, il tire au hasard parmi les tickets de caisse d'une année et obtient :

- 92 clients ont acheté à la fois des bières et des chips,
- 32 clients ont acheté des bières mais pas de chips,
- 10 clients ont acheté des chips mais pas de bières,
- 12 clients nont acheté ni bières ni chips.

Il ne veut se tromper qu'une fois sur 100 en disant qu'il y a un lien entre ces deux types d'achat. Proposer un test statistique adapté au problème.

Exercice 12

Dans une grande entreprise, on a évalué le niveau de stress au travail et mesuré le temps en minutes mis pour se rendre au travail de 550 salariés. Les résultats sont consigné dans le tableau suivant :

Durée trajet (min) / Niveau de stress	< 15	$[15; 45]$	> 45
faible	91	136	48
modéré	39	37	38
élevé	38	69	54

Est-ce que le temps mis pour se rendre au travail a une influence sur le niveau de stress ?