Exercices avec le logiciel R

Exercices auxquels vous avez échappé

Pr Jean R. Lobry

24 février 2017

Cette fiche reprend de vieux problèmes tels qu'ils étaient utilisés en enseignement avant la révolution de R. Les énoncés originaux sont encadrés.

Table des matières

1 Statistiques descriptives

1.1 Moyenne et variance de distances d'insectes

On a mesuré la distance parcourue par un insecte en 30 s. Pour un lot de 10 insectes les résultats ont été les suivants (en mm) :
78 - 170 - 173 - 190 - 90 - 174 - 166 - 293 - 149 - 117
Calculer la moyenne et la variance des distances parcourues.

Cet exercice demandait beaucoup de temps pour calculer à l'aide d'une simple calculatrice la moyenne et la variance de l'échantillon. Il y avait des recettes calculatoires à connaître pour minimiser le nombre d'opérations à effectuer. Tout ceci est devenu complètement obsolète avec R :

Entrez une fois pour toutes les données sous la forme d'un vecteur :

```
dists <- c(78, 170, 173, 190, 90, 174, 166, 293, 149, 117)
```

Calculer la moyenne :

```
mean(dists)
```
```
[1] 160
```

Calculer la variance, c'est à dire la moyenne des carrés des écarts à la moyenne :

```
mean((dists-mean(dists))^2)
```
```
[1] 3268.4
```

On faisait plutôt utiliser aux étudiants la formule suivante :

```
mean(dists^2) -mean(dists)^2
```
```
[1] 3268.4
```

Il y a bien entendu une fonction pré-définie dans R pour calculer la variance, mais celle ci donne l'estimation de la variance de la population, et non la variance de l'échantillon :

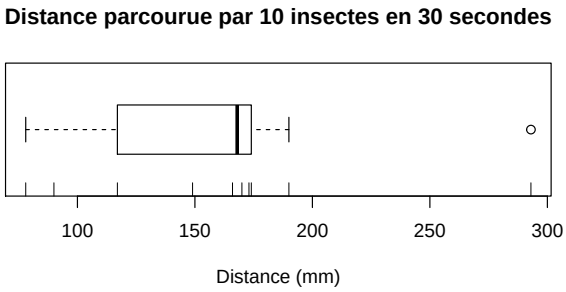
```
var(dists)
```
```
[1] 3631.556
```

Il faut corriger par $\frac{n-1}{n}$ pour retrouver la variance de l'échantillon :

```
var(dists)*9/10
```
```
[1] 3268.4
```
```
var.n <- function(x) var(x)*(length(x)-1)/length(x)
var.n(dists)
```
```
[1] 3268.4
```

En fait, pour apprécier la variabilité de l'échantillon, on commencerait maintenant par une représentation graphique, par exemple :

```
boxplot(dists, horizontal=T, xlab = "Distance (mm)",
main = "Distance parcourue par 10 insectes en 30 secondes")
rug(dists,0.1)
```



On voit tout de suite qu'il y a un individu un peu particulier.

1.2 Taille de 40 élèves

On a mesuré la taille (en cm) de 40 élèves d'une classe. Les résultats sont les suivants :
138 164 150 132 144 125 149 157 146 158 140 147 136 148 152 144 168 126 138 176 163 119 154 165 146 173 142 147 135 153 140 135 161 145 135 142 150 156 145 128
a) Calculer la moyenne des tailles.
b) Regrouper les données en 10 classes puis en 5 classes seulement. Représenter graphiquement les résultats dans les deux cas. Calculer la moyenne dans les deux cas.

Cet exercice était particulièrement fastidieux à faire à la main. Pour se faciliter la tâche il fallait commencer par trier les données, et à la main c'est très pénible. Ensuite il fallait définir les limites des classes, puis attribuer les individus à chaque classe. Quand un individu tombait sur une limite de classe il fallait faire attention si on avait décidé de borner de façon inclusive à droite ou bien à gauche les intervalles correspondants aux classes. Il était très facile de faire une erreur d'inattention (ce n'est pas aussi automatique que de calculer des sommes de valeurs comme pour le calcul de la moyenne et de la variance).

Dans mon souvenir, les étudiants étaient complètement dégoûtés après avoir regroupé les données en 10 classes, et n'avaient plus aucune envie de faire le même exercice en 5 classes (même si c'est trivial une fois que l'on fait le premier regroupement). Et l'objectif pédagogique de montrer l'arbitraire et la perte d'information du regroupement en classes était largement perdu pour la majorité. C'est pourtant un point fondamental que de savoir qu'une ré-écriture des données n'est pas innocente et qu'il y a des compromis à trouver.

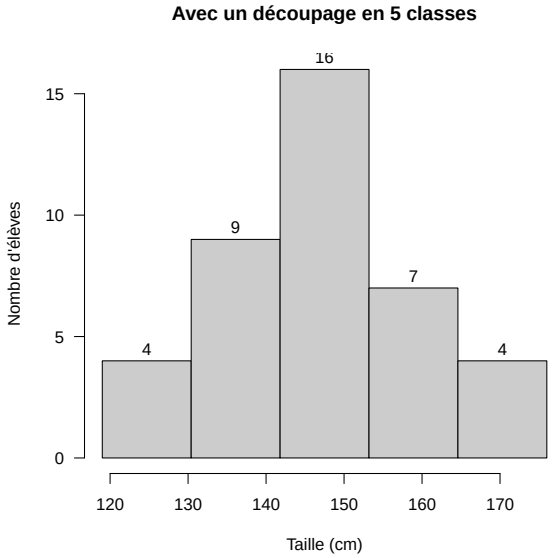
Maintenant sous R on peut peut être aborder ces questions plus facilement. La question a) est directe et sans intérêt :

```
elev <- c(
138, 164, 150, 132, 144, 125, 149, 157,
146, 158, 140, 147, 136, 148, 152, 144,
168, 126, 138, 176, 163, 119, 154, 165,
146, 173, 142, 147, 135, 153, 140, 135,
161, 145, 135, 142, 150, 156, 145, 128)
mean(elev)
```
```
[1] 146.8
```

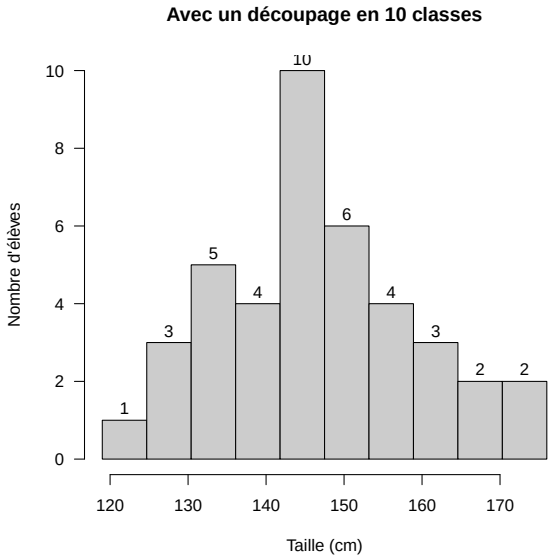
Le b) est plus intéressant : que se passe-t-il lorsque l'on regroupe les données en classes puis que l'on calcule la moyenne ? Il y a perte d'information, c'était le

but de cet exercice. C'est un peu difficile à faire sous R maintenant parce que c'est un problème qui ne se pose plus vraiment sous cette forme. Mais essayons quand même.

```
hist(elev, breaks = seq(from = min(elev), to = max(elev), length = 6), col = grey(0.8),
labels = TRUE, las = 1, xlab = "Taille (cm)", ylab = "Nombre d'élèves",
main = "Avec un découpage en 5 classes") -> avec5
```




```
hist(elev, breaks = seq(from = min(elev), to = max(elev), length = 11), col = grey(0.8),
labels = TRUE, las = 1, xlab = "Taille (cm)", ylab = "Nombre d'élèves",
main = "Avec un découpage en 10 classes") -> avec10
```



Bien, c'est clair et immédiat, il est clair que l'on ne voit pas du tout la même chose avec le découpage en 5 classes et 10 classes. On devine un certain dimorphisme sexuel avec le découpage en 10 classes qui est invisible avec le découpage en 5 classes. Au niveau du calcul des moyennes, voyons ce qu'il en est :

```
(moy5 <- weighted.mean(avec5$mids, avec5$counts))
```

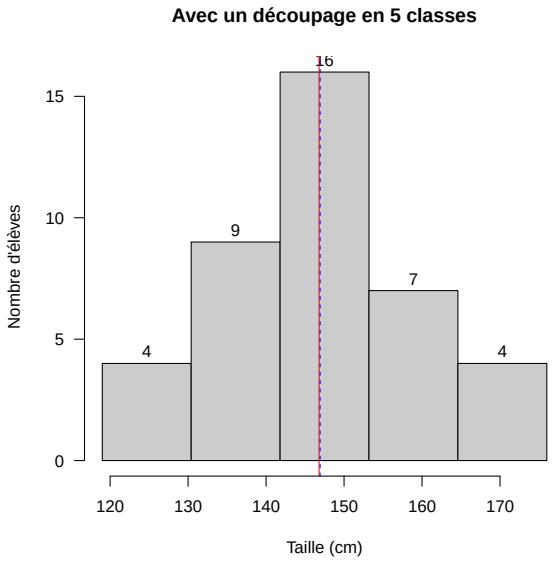
```
[1] 146.93
```

```
(moy10 <- weighted.mean(avec10$mids, avec10$counts))
```

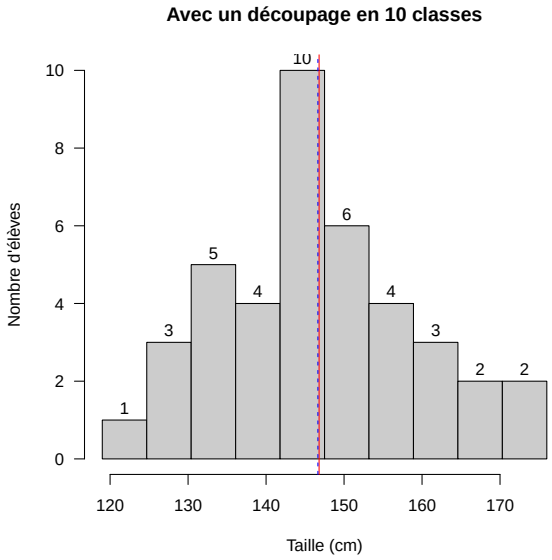
```
[1] 146.645
```

L'information est dégradée, mais visualisons cela :

```
hist(elev, breaks = seq(from = min(elev), to = max(elev), length = 6), col = grey(0.8),
labels = TRUE, las = 1, xlab = "Taille (cm)", ylab = "Nombre d'élèves",
main = "Avec un découpage en 5 classes") -> avec5
abline(v=mean(elev), col="red")
abline(v = moy5, col="blue", lty=2)
```

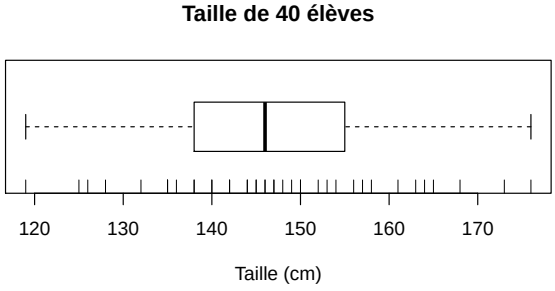



```
hist(elev, breaks = seq(from = min(elev), to = max(elev), length = 11), col = grey(0.8),
labels = TRUE, las = 1, xlab = "Taille (cm)", ylab = "Nombre d'élèves",
main = "Avec un découpage en 10 classes") -> avec10
abline(v=mean(elev), col="red")
abline(v = moy10, col="blue", lty=2)
```



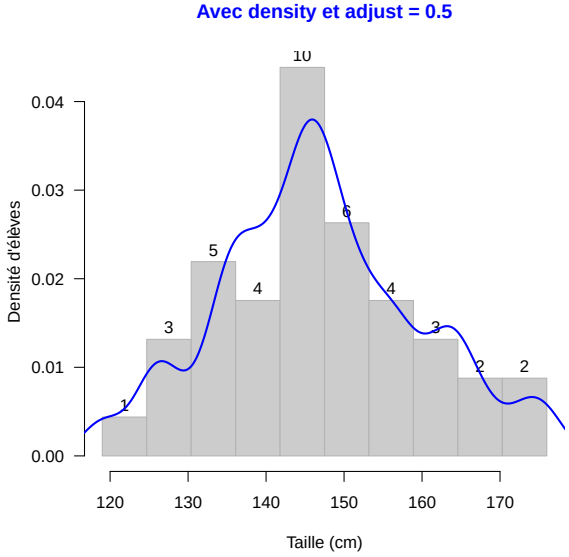
Très joli, la moyenne est ici très peu sensible au choix du regroupement par classes. Mais les indicateurs scalaires de tendance centrale perdent un peu de leur intérêt devant la facilité de mise en ouvre des outils de description de la variabilité. Allons y pour le premier réflexe sous R face à ces données :

```
boxplot(elev, horizontal=T, xlab = "Taille (cm)", main = "Taille de 40 élèves")
rug(elev,0.1)
```



On peut aussi utiliser des estimateurs de la densité locale :

```
adj <- 0.5
brk <- seq(from = min(elev), to = max(elev), length = 11)
dst <- density(x = elev, adjust = adj)
hst <- hist(x = elev, breaks = brk, plot = FALSE )
hist(x = elev, breaks = brk,
ylim = c(0, max(hst$density, dst$y)),
col = grey(0.8),
border = grey(0.7),
labels = as.character(hst$count),
las = 1,
xlab = "Taille (cm)", ylab = "Densité d'élèves",
main = paste("Avec density et adjust =", adj),
probability = TRUE,
col.main = "blue")
lines(dst, lwd = 2, col = "blue")
```

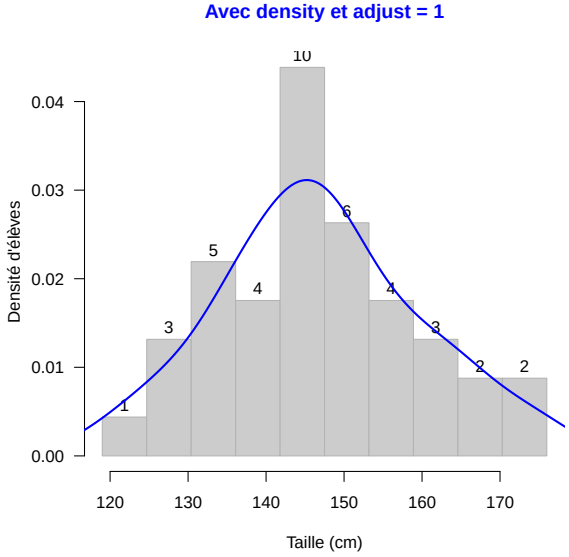


Voir aussi « Statistique pour historiens » par Alain Guerreau :

`http://elec.enc.sorbonne.fr/statistiques/stat2004.pdf`

pour une justification de l'exploration du paramètre `adjust`.

```
adj <- 1
dst <- density(x = elev, adjust = adj)
hist(x = elev, breaks = brk,
ylim = c(0, max(hst$density, dst$y)),
col = grey(0.8),
border = grey(0.7),
labels = as.character(hst$count),
las = 1,
xlab = "Taille (cm)", ylab = "Densité d'élèves",
main = paste("Avec density et adjust =", adj),
probability = TRUE,
col.main = "blue")
lines(dst, lwd = 2, col = "blue")
```



Le gros avantage avec R c'est que l'on peut maintenant travailler avec des données prélevées *in situ*, voir par exemple le TD Variables Estudiantines `https://pbil.univ-lyon1.fr/R/fichestd/tdr314.pdf`.

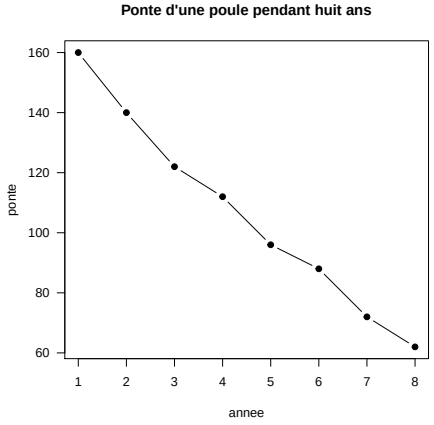
1.3 Ponte d'une poule pendant huit ans

Le nombre d'oeufs pondus en un an par une poule a été relevé pendant 8 ans. Les résultats sont les suivants :

Année (t)	1	2	3	4	5	6	7	8
Nombre d'oeufs pondus (n)	160	140	122	112	96	88	72	62

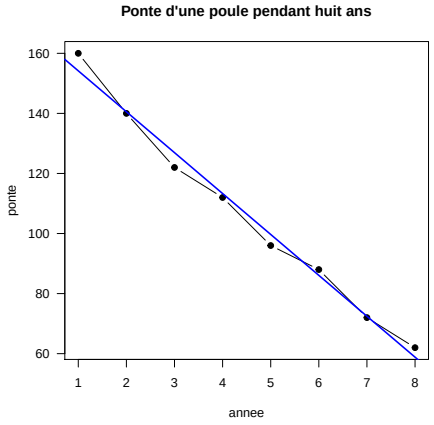
a) Représenter graphiquement les données. Quelle relation entre n et t vous suggère cette représentation ? b) Calculer la moyenne et la variance du nombre d'oeufs pondus.

```
annee <- 1:8
ponte <- c(160 ,140 ,122 ,112 ,96 ,88 ,72 ,62)
plot(x = annee, y = ponte, type = "b",
main = "Ponte d'une poule pendant huit ans", las = 1, pch = 19)
```



Cela suggère une décroissance linéaire des pontes au cours du temps :

```
annee <- 1:8
ponte <- c(160 ,140 ,122 ,112 ,96 ,88 ,72 ,62)
plot(x = annee, y = ponte, type = "b",
main = "Ponte d'une poule pendant huit ans", las = 1, pch = 19)
abline(lm(ponte~annee), col = "blue",lwd=2)
```



Calcul de la moyenne et de la variance de l'échantillon :

```
 mean(ponte)
[1] 106.5
 mean((ponte - mean(ponte))^2)
[1] 984.75
 var(ponte)
[1] 1125.429
```

Je ne vois pas trop l'intérêt de cet exercice à part vérifier que les étudiants aient bien compris ce que sont des données regroupées (ici les données ne sont pas regroupées, contrairement à ce que pourrait suggérer la présentation).

1.4 Accidents à Pearson-Gulch

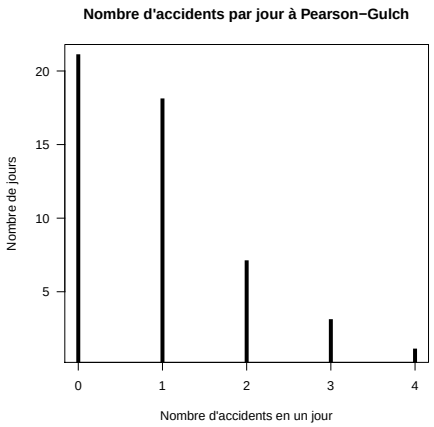
On a enregistré, pendant 50 jours, le nombre d'accidents automobiles à PEARSON-GULCH. Les résultats sont les suivants :

Nombre d'accidents x_i	0	1	2	3	4
Nombre de jours (n_i) où il s'est produit x_i accidents	21	18	7	3	1

a) Représenter graphiquement les données. b) Calculer la moyenne et la variance du nombre d'accidents. c) Quelle loi de probabilité pourrait suivre ce nombre d'accidents par jour ?

Un grand classique, calculer la moyenne et la variance sur des données regroupées, facile ici puisqu'ici on donne directement les n_i et les x_i, donc il n'y a pas de confusion possible à ce niveau. Puis, surprise, la moyenne et la variance sont du même ordre de grandeur, ce qui suggère une loi de Poisson. À propos de la loi de Poisson, je me souviens que Daniel Sillans s'est retrouvé par le hasard des choses à avoir à introduire la loi de Poisson en amphi un jour de premier avril. Les étudiants n'ont jamais voulu le croire.

```
x <- c(0 ,1 ,2 ,3 ,4 )
n <- c(21 ,18 ,7 ,3 ,1)
par(lend = "square")
my <- "Nombre de jours"
mx <- "Nombre d'accidents en un jour"
ma <- "Nombre d'accidents par jour à Pearson-Gulch"
plot(x = x, y = n, type = "h", lwd = 5, las = 1,main = ma,ylab = my,xlab = mx)
```



Toute la difficulté ici vient de ce que les données sont regroupées. Il est facile sous R de reconstituer les données originelles :

```
orig <- rep(x,n)
```

Donc à partir de là il est facile de calculer la moyenne et la variance de l'échantillon :

```
mean(orig)
[1] 0.9
mean((orig-mean(orig))^2)
[1] 0.97
```

Les valeurs de la moyenne et de la variance sont proches, ce qui suggère une loi de Poisson. Est-ce qu'il y a moyen sous R de calculer la moyenne et la variance à partir des données regroupées ? Pour la moyenne cela ne pose pas de problème puisqu'il y a la fonction pré-définie `weighted.mean()` pour calculer des moyenne pondérées :

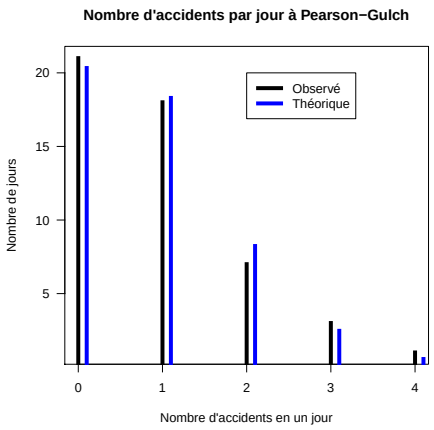
```
weighted.mean(x =x, w = n)
```
```
[1] 0.9
```

Je n'ai pas trouvé l'équivalent pour la variance (ça doit bien pourtant exister quelque part, grrrr). Bon, ce n'est pas grave,

```
weighted.mean((x-weighted.mean(x, n))^2, n)
```
```
[1] 0.97
```

On aimerait bien savoir s'il est raisonnable de penser que l'on a bien une loi de Poisson ici. Représentons ce qui serait attendu sous une loi de Poisson de paramètre λ égal à la moyenne de notre échantillon :

```
x <- c(0 ,1 ,2 ,3 ,4 )
n <- c(21 ,18 ,7 ,3 ,1)
par(lend = "square")
my <- "Nombre de jours"
mx <- "Nombre d'accidents en un jour"
ma <- "Nombre d'accidents par jour à Pearson-Gulch"
plot(x = x, y = n, type = "h", lwd = 5, las = 1,main = ma,ylab = my,xlab = mx)
lines(x = x + 0.1, y = dpois(0:4,mean(orig))*sum(n), type = "h", col = "blue", lwd = 5)
legend(2, 20, c("Observé","Théorique"), col = c("black","blue"), lwd = 5)
```



Comme c'est joli ! Mais on peut avoir oublié que l'estimateur au maximum de vraisemblance du λ de la loi de Poisson est donné par la moyenne de l'échantillon. Ce n'est pas grave, utilisons l'artillerie lourde :

```
library(MASS)
orig
```
```
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2
[41] 2 2 2 2 2 2 3 3 3 4
```
```
mean(orig)
```
```
[1] 0.9
```
```
fitdistr(orig, dpois, list(lambda=1))
```

```
    lambda
 0.9000000
(0.1341639)
```

C'est un exemple issu de la vie réelle :

```
From: Prof Brian Ripley <ripley>
Date: Thu, 25 Aug 2005 08:56:01 +0100 (BST)

On Thu, 25 Aug 2005, Mark Miller wrote:

> I am trying to fit a number of distributions to a set of data, I have used the
> fitdristr() funtion for most of the distributions, but Poisson is not one of
> the possible distributions. I found somewhee talking about using the gamlss
> package, but have been unable to find it again, any help would be greatly
> appreciated.

If you mean fitdistr in package MASS, it can be used. But the mle for a
Poisson is just the sample mean.


> x <- rpois(250, 2.6)
> mean(x)
[1] 2.62
> fitdistr(x, dpois, list(lambda=2))
     lambda
   2.6203125
  (0.1023841)

--
Brian D. Ripley,                  ripley at stats.ox.ac.uk
Professor of Applied Statistics,  http://www.stats.ox.ac.uk/~ripley/
University of Oxford,             Tel:  +44 1865 272861 (self)
1 South Parks Road,                     +44 1865 272866 (PA)
Oxford OX1 3TG, UK                Fax:  +44 1865 272595
```

La question sous-jacente dans cet exercice est de savoir s'il est raisonnable de penser que la loi du nombre d'accidents journaliers à Pearson-Gulch suit une loi de Poisson de paramètre $\lambda \approx 0.9$. Rien de plus facile maintenant sous R. On calcule les effectifs théoriques sous le modèle :

```
theo <- dpois(0:4,mean(orig))
sum(theo)
```

```
[1] 0.9976559
```

On colle la queue de la distribution dans la dernière classe :

```
theo[5] <- 1 - sum(theo[1:4])
sum(theo)
```

```
[1] 1
```

On fait un test du χ^2 d'ajustement à une distribution :

```
(chisq.test(x = n, p = theo) -> testchi)
```

```
        Chi-squared test for given probabilities
data:  n
X-squared = 0.48436, df = 4, p-value = 0.975
```

On a plusieurs problèmes. Le premier est juste un message d'avis disant :

```
Chi-squared approximation may be incorrect in: chisq.test(x = n, p = theo)
```

Ceci est facilement corrigé par :

```
(chisq.test(x = n, p = theo, simulate.p.value = TRUE) -> testchi)
```

Le deuxième problème est sur le calcul des degrés de liberté pour le test du χ^2. On a estimé un paramètre, donc la vraie p-value est :

```
1-pchisq(testchi$statistic, df = 3)
```

```
X-squared
0.9223147
```

Ce qui ne change pas grand chose, avec un risque de première espèce de 5 % les données ne permettent pas de rejeter l'hypothèse nulle d'une loi de Poisson de paramètre $\lambda = 0.9$.

1.5 Cécidomyie du hêtre

La cécidomyie du hêtre provoque sur les feuilles de cet arbre des galles dont la distribution a été observée : x est le nombre de galles par feuille, n est le nombre de feuilles portant x galles :

x	0	1	2	3	4	5	6	7	8	9	10
n	482	133	46	24	6	5	2	1	0	1	0

a) Représenter graphiquement les données. b) Calculer la moyenne et la variance du nombre de galles par feuille.

Cet énoncé m'évoque tout de suite une vieille blague : Deux enfants de six ans discutent ensemble.... L'un dit à l'autre : - Eh, ce matin, j'ai trouvé une ...[1] dans la véranda. Et l'autre lui répond : - C'est quoi une véranda ?

Avant de faire des représentations graphiques, essayons de nous approprier un minimum les données. D'après la structure grammaticale de l'énoncé, il semble que « hêtre » soit une instance de la classe « arbre ». Un coup de moteur de recherche internet sur « hêtre » dans les images nous donne :

`http://www.tarn-web.com/cards/hetre.jpg`

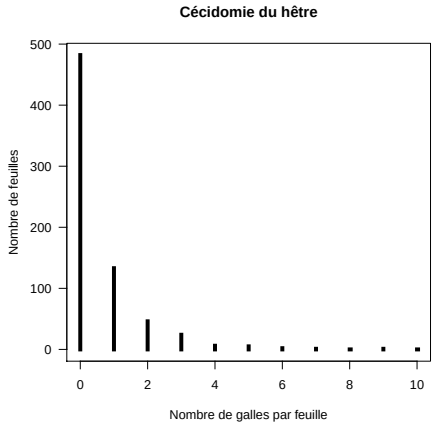
Donc un « hêtre » semble bien être un arbre. Toujours d'après l'énoncé, il peut avoir des « galles », un coup de google sur « hêtre galle » dans les images nous donne :

1. Vous pouvez choisir ce paramètre suivant votre sensibilité, par exemple *Cataglyphis piliscapus*

http://aramel.free.fr/INSECTES40.shtml

C'est un truc pas sympathique pour les feuilles d'un hêtre. Maintenant on comprend le sens de l'énoncé : on a compté le nombre de ces verrues fort peu esthétiques sur les feuilles d'un arbre d'une espèce bien identifiée (un hêtre). Reste « cécidomyie », une petite recherche montre l'ampleur du problème. Il n'y a donc pas que le hêtre qui soit concerné.

```
x <- 0:10
n <- c(482 ,133 ,46 ,24 ,6 ,5 ,2 ,1 ,0 ,1 ,0)
par(lend = "square")
plot(x = x, y = n, type = "h", lwd = 5, las = 1,
main = "Cécidomie du hêtre",
xlab = "Nombre de galles par feuille",
ylab = "Nombre de feuilles")
```



Calculons la moyenne et la variance :

```
mean(rep(x,n))
[1] 0.5342857
var(rep(x,n))
[1] 1.081798
```

1.6 Levures dans un hématimètre

Dans les 400 carrés d'un "hématimètre", on a compté 1872 levures, avec la répartition suivante : x est le nombre de levures par carré, n est le nombre de carrés.

x	0 ou 1	2	3	4	5	6	7	8	9	10 et plus
n	20	43	53	86	70	54	37	18	10	9

Calculer le nombre moyen de levures par carré. Quelle loi pourrait suivre le nombre de levures par carré ?

```
x <- c(0.5, 2:10)
n <- c(20 ,43 ,53 ,86 ,70 ,54 ,37 ,18 ,10 ,9)
sum(n)
```

```
[1] 400
```

```
mean(rep(x,n))
```

```
[1] 4.64
```

```
var(rep(x,n))
```

```
[1] 4.479098
```

Cela suggère donc une loi de Poisson de paramètre $\lambda = 4.64$.

```
(dpois(0:10, lambda = mean(rep(x,n)))->tmp)
```

```
 [1] 0.009657698 0.044811717 0.103963183 0.160796390 0.186523813 0.173094098
 [7] 0.133859436 0.088729683 0.051463216 0.026532147 0.012310916
```

```
sum(tmp)
```

```
[1] 0.9917423
```

```
tmp[11] <- 1 - sum(tmp[1:10])
sum(tmp)
```

```
[1] 1
```

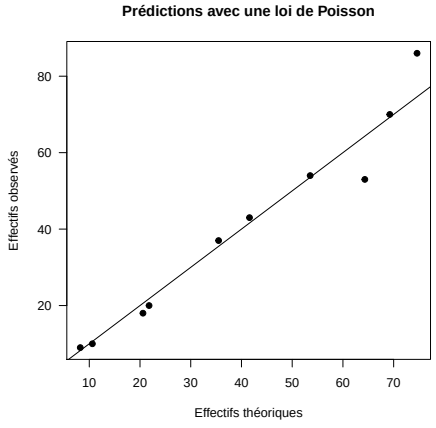
```
tmp[2] <- tmp[2]+tmp[1]
tmp <- tmp[-1]
sum(tmp)
```

```
[1] 1
```

```
(theo <- 400*tmp)
```

```
 [1] 21.787766 41.585273 64.318556 74.609525 69.237639 53.543774 35.491873 20.585287
 [9] 10.612859  8.227447
```

```
plot(theo, n, pch = 19, las = 1,
xlab ="Effectifs théoriques",
ylab ="Effectifs observés",
main ="Prédictions avec une loi de Poisson")
abline(c(0,1))
```



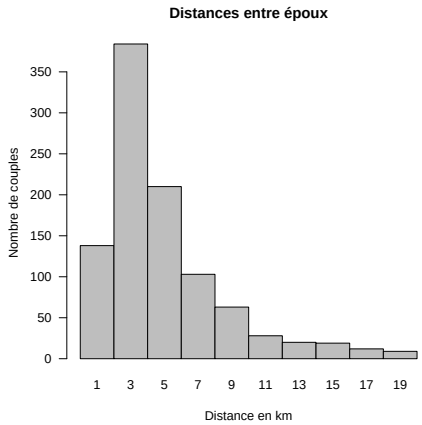
1.7 Distances entre époux

Une enquête concernant les distances entre domiciles des époux, au moment de leur mariage, a donné, dans le Finistère, les résultats suivants :

Distance en km	Nombre de couples
0 - 2	138
2 - 4	384
4 - 6	210
6 - 8	103
8 - 10	63
10 - 12	28
12 - 14	20
14 - 16	19
16 - 18	12
18 - 20	9

a) Représenter graphiquement les données. b) Calculer la moyenne et la variance des distances.

```
x <- 2*(0:9)+1
n <- c(138,384,210,103,63,28,20,19,12,9)
barplot(n,names.arg=x,space=0,
 las = 1,
xlab ="Distance en km",
ylab ="Nombre de couples",
main = "Distances entre époux")
```




```
mean(rep(x,n))
```
```
[1] 4.924949
```
```
var(rep(x,n))
```
```
[1] 13.03396
```

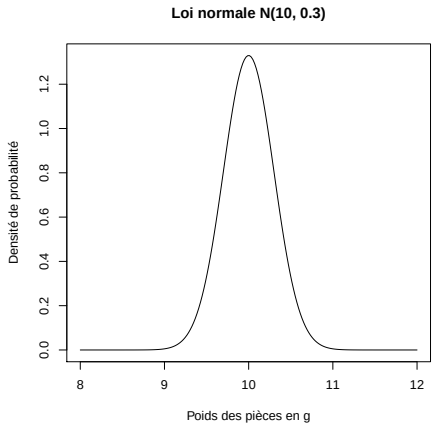
2 Intervalle de confiance de moyennes

2.1 Pièces métalliques

Dans une entreprise on fabrique des pièces métalliques. Le poids moyen de ces pièces est de 10 g. Les calculs faits sur des statistiques antérieures ont montré que l'écart-type du poids de ces pièces est de 0,3 g.
Au cours d'une certaine période on fabrique un très grand nombre de ces pièces. Comme il n'est pas possible de contrôler que toutes les pièces pèsent en moyenne 10 g, on en pèse 100, prises au hasard dans la production de la période considérée. Le poids moyen de ces 100 pièces est $\bar{x} = 10{,}05$ g.
a) Que penser de ce résultat ?
b) Est-il en contradiction avec la moyenne de 10 g jusque-là admise ?

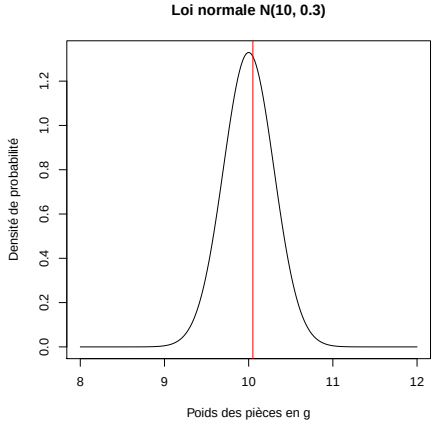
On va essayer de répondre sans faire appel à la moindre formule. Voyons la loi théorique du poids des pièces.

```
poids <- seq(from = 8, to = 12, length = 200)
prob <- dnorm(poids, mean = 10, sd = 0.3)
plot(poids,prob, type = "l",
xlab = "Poids des pièces en g",
ylab ="Densité de probabilité",
main="Loi normale N(10, 0.3)")
```



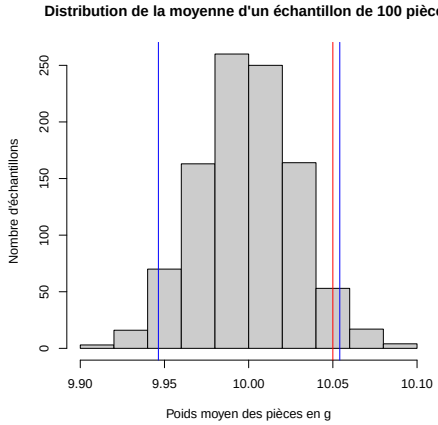
La valeur moyenne observée sur un échantillon de 100 pièces ne semble pas être aberrante :

```
poids <- seq(from = 8, to = 12, length = 200)
prob <- dnorm(poids, mean = 10, sd = 0.3)
plot(poids,prob, type = "l",
xlab = "Poids des pièces en g",
ylab ="Densité de probabilité",
main="Loi normale N(10, 0.3)")
abline(v=10.05, col = "red")
```



On a oublié la formule donnant la loi de la moyenne, donc on simule sous le modèle le tirage d'un grand nombre d'échantillons de 100 pièces :

```
sim <- sapply(1:1000, function(x)mean(rnorm(100,mean=10,sd=0.3)))
hist(sim, main = "Distribution de la moyenne d'un échantillon de 100 pièces",
xlab = "Poids moyen des pièces en g",
ylab ="Nombre d'échantillons",
col = grey(0.8))
abline(v=10.05, col = "red")
abline(v=quantile(sim, probs = c(0.025,0.975)), col = "blue")
```



On reste juste dans les limites du raisonnable puisque dans 95 % des cas on tombe entre les bornes en bleu. Notre échantillon à 10.05 g n'a rien d'extraordinaire. Et pour faire un vrai test, avec un tableau de probabilités critiques :

```
a1 <- round(1-pnorm((10.05-10)*sqrt(100)/0.3),dig=3)
a2 <- round(pnorm((10.05-10)*sqrt(100)/0.3),dig=3)
a3 <- 2*min(a1,a2)
```

Trop grande	Trop petite	Différente
0.048	0.952	0.096

On est vraiment jute dans les limites du raisonnable. L'acheteur et le vendeur ne prendrons peut-être pas la même décision et on refera quelques mesures.

2.2 Risques de première et de seconde espèces

Le but du problème est de tester l'hypothèse qu'une pièce de monnaie n'est pas "truquée". La règle suivante de décision est adoptée :
Acceptation de l'hypothèse si le nombre de faces obtenu en lançant 100 fois la pièce appartient à l'intervalle [40, 60].
Rejet de l'hypothèse dans la cas contraire.
a) Quel est le risque de première espèce, c'est-à-dire quelle est la probabilité de rejeter l'hypothèse alors qu'elle est vraie ?
b) Interpréter graphiquement la règle de décision adoptée et le résultat de la question précédente.
c) Quelle conclusion tirez-vous si vous obtenez 53 fois "face" lors de 100 lancers ? 60 fois "face" ?
d) Pouvez-vous vous tromper dans vos conclusions à la question précédente ? Quelle est la probabilité d'accepter l'hypothèse que la pièce n'est pas truquée alors que l'on sait par ailleurs que la probabilité réelle d'obtenir "face" est 0,6 ? (risque de seconde espèce).
e) Calculer à nouveau le risque de seconde espèce ? lorsque la probabilité réelle d'obtenir "face" est p = 0,7 ; 0,8 ; 0,9 ; 0,4.
f) Représenter graphiquement β et 1 - β en fonction de p. Conclusions

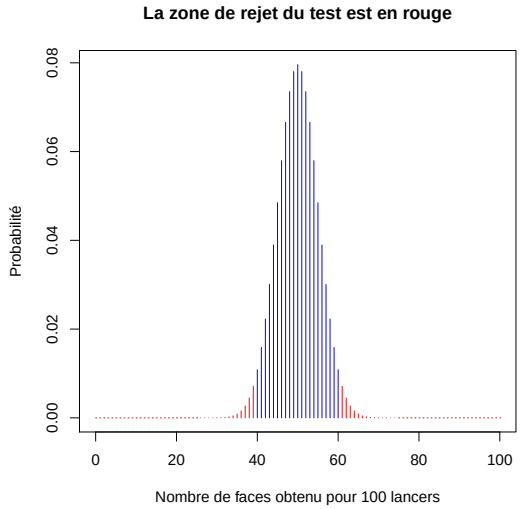
Le a) est direct sous R :

```
1-sum(dbinom(x = 40:60,size = 100, prob = 0.5))
```

```
[1] 0.0352002
```

Ce qui correspond à la zone de rejet en rouge :

```
plot(x = 0:100, y = dbinom(x = 0:100,size = 100, prob = 0.5), type = "h",
col = rep(c("red","blue","red"), c(40, 21, 40)), xlab = "Nombre de faces obtenu pour 100 lancers",
ylab = "Probabilité",
main = "La zone de rejet du test est en rouge")
```



Pour le c) on peut écrire le test :

```
montest <- function(x)
{
  if(x >= 40 & x <= 60 )
    return("Acceptation Ho")
  else
    return("Rejet Ho")
}
 montest(53)
```

```
[1] "Acceptation Ho"
```

```
 montest(60)
```

```
[1] "Acceptation Ho"
```

Le d) est direct :

```
sum(dbinom(x = 40:60,size = 100, prob = 0.5))
```

```
[1] 0.9647998
```

Le e) aussi :

```
sum(dbinom(x = 40:60,size = 100, prob = 0.7))
```

```
[1] 0.02098858
```

```
sum(dbinom(x = 40:60,size = 100, prob = 0.8))
```

```
[1] 3.60842e-06
```

```
sum(dbinom(x = 40:60,size = 100, prob = 0.9))
```

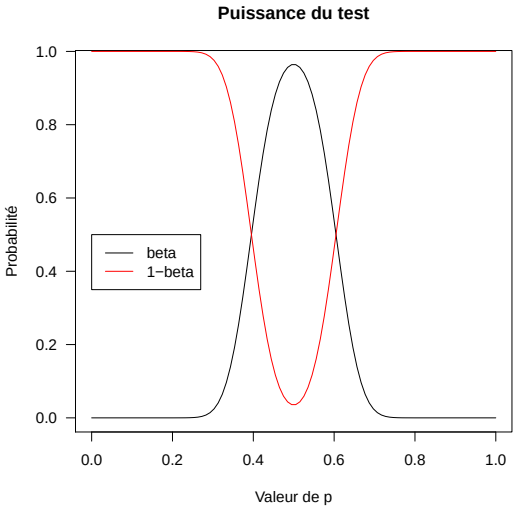
```
[1] 2.94553e-15
```

```
sum(dbinom(x = 40:60,size = 100, prob = 0.4))
```

```
[1] 0.5379066
```

Le f) montre que le test est d'autant plus puissant que la pièce est biaisée :

```
ps <- seq(from=0, to = 1, length = 100)
f <- function(x) sum(dbinom(x = 40:60,size = 100, prob = x))
beta <- sapply(ps, f)
plot(ps, beta, type = "l", col = "black", las = 1,
xlab = "Valeur de p",
ylab = "Probabilité",
main = "Puissance du test")
lines(ps, 1-beta, col="red")
legend(0,0.5,c("beta","1-beta"),col=c("black","red"),lwd=1)
```



2.3 Balance de laboratoire

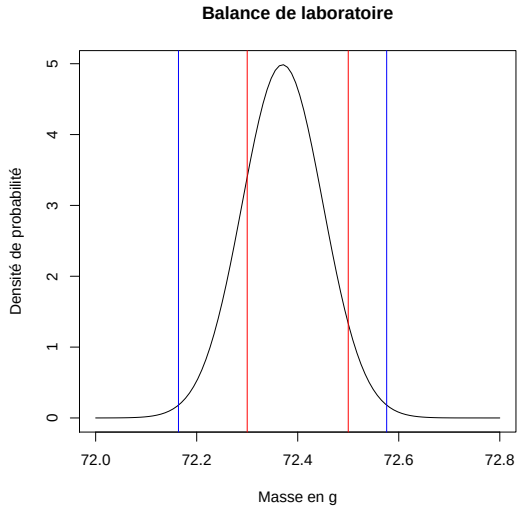
On admet qu'avec la balance utilisée dans un laboratoire, le résultat de la pesée d'un corps de masse comprise entre 50 et 100 g est une variable aléatoire ayant une distribution normale N $(\mu, 0,08)$. Dans le cas où μ = 72,37 g :
Calculer la probabilité de l'événement $72,30 < X < 72,50$
Déterminer un intervalle [μ - h, μ + h] tel que la probabilité que X prenne une valeur dans cet intervalle soit 0,99.

```
pnorm(q = 72.5,mean=72.37,sd=0.08)-pnorm(q = 72.3,mean=72.37,sd=0.08)
```
```
[1] 0.7571318
```

```
qnorm(p=0.005,mean=72.37,sd=0.08)
```
```
[1] 72.16393
```
```
qnorm(p=0.995,mean=72.37,sd=0.08)
```
```
[1] 72.57607
```

Graphiquement :

```
ps <- seq(from = 72, to = 72.8, length = 100)
prob <- dnorm(ps, mean = 72.37, sd = 0.08)
plot(ps, prob, type = "l",
xlab = "Masse en g",
ylab = "Densité de probabilité",
main ="Balance de laboratoire")
abline(v= 72.3, col="red")
abline(v= 72.5, col="red")
abline(v = qnorm(p=0.005,mean=72.37,sd=0.08), col ="blue")
abline(v = qnorm(p=0.995,mean=72.37,sd=0.08), col ="blue")
```



2.4 Masse de 20 cocons

Dans un lot de 500 cocons environ d'une race bivoltine de *Bombyx mori*, on a prélevé 20 cocons au hasard et on les a pesés. Les poids (en g) ainsi mesurés sont les suivants.
0,64 0,65 0,73 0,60 0,65 0,77 0,82 0,64 0,66 0,72 0,87 0,84 0,66 0,76 0,63 0,52 0,66 0,45 0,74 0,79
a) Calculer le poids moyen et la variance de cet échantillon. b) Donner l'intervalle de confiance au seuil 5 % du poids moyen d'un cocon de la population considérée.

```
cocons <- c(0.64, 0.65, 0.73, 0.60, 0.65, 0.77, 0.82, 0.64, 0.66,
0.72,
0.87 ,0.84, 0.66, 0.76, 0.63, 0.52, 0.66, 0.45, 0.74,
0.79 )
mean(cocons)
```

```
[1] 0.69
```

```
mean((cocons-mean(cocons))^2)
```

```
[1] 0.01036
```

```
var(cocons)
```

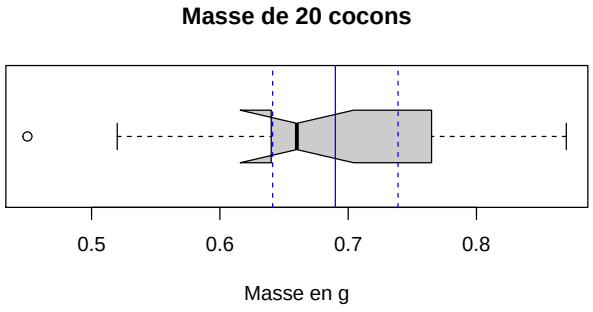
```
[1] 0.01090526
```

On peut comparer l'intervalle de confiance de la moyenne avec celui de la médiane :

```
t.test(cocons)$conf
```

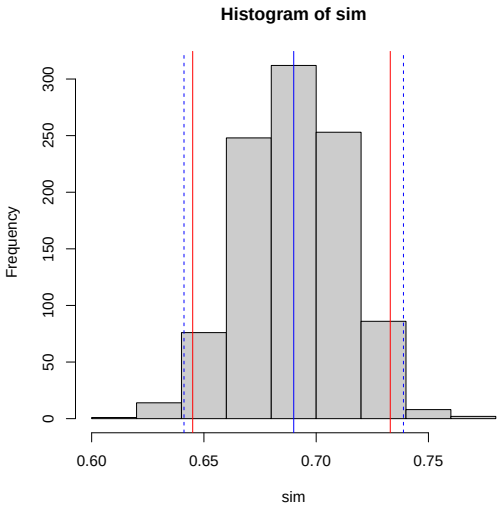
```
[1] 0.6411261 0.7388739
attr(,"conf.level")
[1] 0.95
```

```
boxplot(cocons, notch=TRUE, horizontal = TRUE, col = grey(0.8),
xlab = "Masse en g",
main = "Masse de 20 cocons")
abline(v = mean(cocons), col ="blue")
abline(v = t.test(cocons)$conf, col ="blue", lty = 2)
```



On peut comparer avec une approche par simulation :

```
replicate(1000, mean(sample(cocons, replace = TRUE)))->sim
hist(sim, col = grey(0.8))
abline(v=quantile(sim, probs = c(0.025,0.975)), col = "red")
abline(v = mean(cocons), col ="blue")
abline(v = t.test(cocons)$conf, col ="blue", lty = 2)
```



2.5 Masse d'un corps

Pour des masses comprises entre 50 g et 200 g, la variance du résultat de la pesée est 0,0015. Les résultats des trois pesées d'un même corps ont été : 64,32 ; 64,27 ; 64,39 . Donner l'intervalle de confiance au seuil 5 % du poids moyen de ce corps.

```
 x <- c(64.32 ,   64.27 ,   64.39)
 t.test(x)$conf
[1] 64.17693 64.47640
attr(,"conf.level")
[1] 0.95
```

Sauf que c'est pas bon ici puisque la variance est donnée.

```
 qnorm(p= c(0.025,0.975), mean = mean(x), sd = sqrt(0.0014/3))
[1] 64.28433 64.36901
```

2.6 Glucose sanguin

Des études statistiques montrent que le taux de glucose dans le sang est une variable Gaussienne X de moyenne 1 g/l et d'écart-type 0,1 g/l. On prend un échantillon de 9 individus. Quels sont la moyenne et l'écart-type théorique ($\mu_{\bar{x}}$ et $\sigma_{\bar{x}}$) de la variable aléatoire $\bar{X}$? Quelle est la loi de $\bar{X}$? On désire savoir si un traitement déplace le taux moyen de glucose. La moyenne observée sur un échantillon de neuf individus traités est $\bar{x} = 1{,}06$ g/l. Quelle hypothèse faites-vous ? Faites le test correspondant.

$$\mu_{\bar{x}} = \mu = 1$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.1}{3}$$

$$\bar{X} \sim \mathcal{N}(\mu_{\bar{x}}, \sigma_{\bar{x}})$$

```
qnorm(p= c(0.025,0.975), mean = 1.06, sd = 0.1/3)
```

```
[1] 0.9946679 1.1253321
```

Avec un risque de première espèce de 5 %, les données expérimentales ne permettent pas de mettre en évidence un effet du traitement sur le glucose sanguin.

2.7 Soies de Drosphiles

On a compté le nombre de soies sur le 4ème segment abdominal d'un lot de Drosophiles. Les résultats sont consignés dans le tableau suivant :

Nombre x de soies	Nombre de mouches ayant x soies
16	2
17	5
18	7
19	3
20	3

On demande : a) de calculer la moyenne de cet échantillon, b) de calculer l'intervalle de confiance de cette moyenne.

```
x <- 16:20
n <- c(2,5,7,3,3)
mean(rep(x,n))
```

```
[1] 18
```

```
t.test(rep(x,n))$conf
```

```
[1] 17.43185 18.56815
attr(,"conf.level")
[1] 0.95
```

2.8 Taille de 100 étudiants

Supposons que les tailles de 100 étudiants masculins de l'Université de BROL-NEGE représentent un échantillon aléatoire des tailles des 4 521 étudiants de cette université. Sachant que dans cet échantillon la taille moyenne vaut 1,73 m et l'écart-type 10 cm, déterminer un intervalle de confiance à 0,95 de la taille moyenne des étudiants de l'université.

```
qnorm(p= c(0.025,0.975), mean = 1.73, sd = 0.1/10)
```

```
[1] 1.7104 1.7496
```

3 Intervalle de confiance de proportions

3.1 Urne

Une urne contient des boules rouges et des boules blanches. On tire successivement 128 boules sans remettre les boules après le tirage : 64 boules rouges ont été tirées. Donner l'intervalle de confiance à 0,95 de la proportion des boules rouges dans l'urne.

```
prop.test(64,128)$conf
```

```
[1] 0.4146522 0.5853478
attr(,"conf.level")
[1] 0.95
```

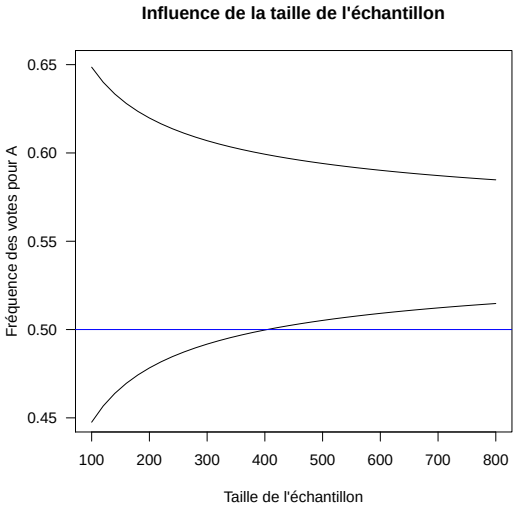
3.2 Votes pour un candidat

Un échantillon de 100 votants choisis au hasard parmi tous les votants d'un bureau donné indique que 55 % d'entre eux votent pour un candidat donné A. a) Donner l'intervalle de confiance à 0,95 de la proportion de tous les votants en faveur du candidat A. b) Quelle aurait du être la taille de l'échantillon pour conclure au niveau 95 % que le candidat A sera élu ?

```
prop.test(55,100)$conf
```

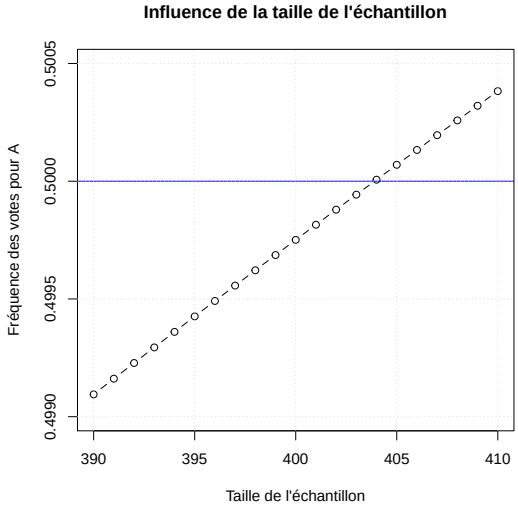
```
[1] 0.4475426 0.6485719
attr(,"conf.level")
[1] 0.95
```

```
n <- seq(from = 100, to = 800, by =20)
f <- function(x) prop.test(0.55*x,x)$conf
tmp <- sapply(n,f)
plot(n,tmp[1,], type = "l", ylim = c(0.45,0.65),
xlab="Taille de l'échantillon", las = 1,
ylab = "Fréquence des votes pour A",
main ="Influence de la taille de l'échantillon")
points(n, tmp[2,], type = "l")
abline(h=0.5, col="blue")
```



Donc il faut compter au moins 400 individus dans l'échantillon. En zoomant un peu :

```
n <- seq(from = 390, to = 410, by =1)
tmp <- sapply(n,f)
plot(n,tmp[1,], type = "b", ylim = c(0.499,0.5005),
xlab="Taille de l'échantillon",
ylab = "Fréquence des votes pour A",
main ="Influence de la taille de l'échantillon")
abline(h=0.5, col="blue")
grid()
```



Il faut au moins 404 individus.

3.3 Pourcentage de fumeurs

On a trouvé que le pourcentage de fumeurs d'une certaine population était de 60 %. On prélève 100 personnes de cette population et on parie que le pourcentage de fumeurs sera compris entre 0,5 et 0,7. Quel est le risque de perdre le pari ?

```
1-sum(dbinom(x = 50:70, size = 100, prob = 0.6))
[1] 0.031537
```

3.4 Formule leucocytaire

Une formule leucocytaire, déterminée par l'examen de 500 éléments blancs, a donné comme pourcentage de lymphocytes 30 %. Peut-on considérer que ce pourcentage diffère significativement de 25 % ?

```
prop.test(x = 0.3*500, 500)$conf
[1] 0.2605275 0.3425979
attr(,"conf.level")
[1] 0.95
```

Oui, on peut considérer que ce pourcentage diffère significativement de 25 %.

3.5 Couleur des yeux de Drosophiles

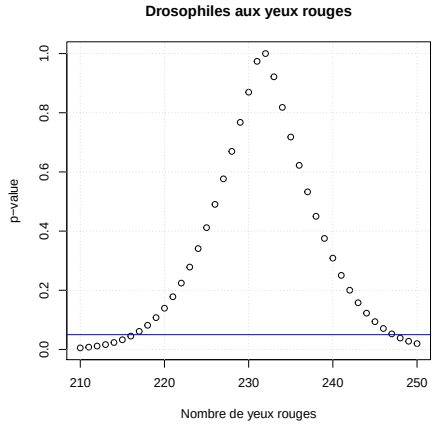
Sur des Drosophiles, on considère le caractère "couleur des yeux". Par croisement d'homozygotes "yeux rouges" AA et d'homozygotes "yeux bruns" aa, on obtient des hétérozygotes Aa. Si on croise ces hétérozygotes entre eux, on doit obtenir (loi de Mendel) 3/4 de "yeux rouges" (AA et Aa), et 1/4 de "yeux bruns" (aa). On a observé, chez 309 drosophiles de cette seconde génération, 241 "yeux rouges" et 68 "yeux bruns". Cette répartition est-elle conforme à la loi de Mendel ? Combien aurait-il fallu observer de drosophiles "yeux rouges", pour décider de rejeter la conformité avec la loi de Mendel, au risque 5 % ?

```
prop.test(x=241,n = 309, p=3/4)
```

```
        1-sample proportions test with continuity correction
data:  241 out of 309, null probability 3/4
X-squared = 1.3215, df = 1, p-value = 0.2503
alternative hypothesis: true p is not equal to 0.75
95 percent confidence interval:
 0.7287441 0.8240130
sample estimates:
        p
0.7799353
```

Les données ne permettent pas de rejeter la loi de Mendel.

```
ns <- 210:250
sapply(ns, function(x) prop.test(x = x, n=309,p=3/4)$p.value)->tmp
plot(ns,tmp,
xlab="Nombre de yeux rouges",
ylab ="p-value",
main="Drosophiles aux yeux rouges")
abline(h=0.05, col="blue")
grid()
```



Si on avait observé moins de 216 (inclus) ou plus de 248 (inclus) Drosophiles aux yeux rouges on aurait rejeté l'hypothèses nulle.

4 Comparaison de moyennes

4.1 Initiation aux tests statistiques

Trois étudiants A B C manipulent en trinôme. On fait l'hypothèse qu'ils ont même adresse. À la fin du stage ils ont cassé 4 boites de Pétri. Quelles sont les probabilités que : a) A ait cassé 3 boites et B une ; b) que l'un ait cassé 3 boites ; c) que l'un ait cassé 4 boites ; d) le fait que 3 boites aient été cassées par le même étudiant conduit-il à penser qu'il est plus maladroit que les autres ?

Remarques : certain modèles de boîtes de Petri sont en verre et peuvent effectivement casser ; elles doivent leur nom à Richard Petri (1852-1921), sans accent sur le e.

```
dmultinom(x=c(3,1,0), size=4, prob=rep(1/3,3))
```

```
[1] 0.04938272
```

```
6*dmultinom(x=c(3,1,0), size=4, prob=rep(1/3,3))
```

```
[1] 0.2962963
```

```
3*dmultinom(x=c(4,0,0), size=4, prob=rep(1/3,3))
```

```
[1] 0.03703704
```

4.2 Compétition larvaire

On se propose d'étudier l'influence de l'intensité de la compétition larvaire entre individus sur le poids des adultes qui émergeront (étude chez un Diptère, la Cératite). Deux lots de larves de Cératite sont repiqués sur milieu artificiel, dans des conditions identiques. Lot 1 : 300 larves, Lot 2 : 600 larves. Du lot n° 1 émergent 250 adultes et du lot n° 2, 450. Ces adultes sont alors anesthésiés et pesés. Soit p la variable poids, exprimée en mg. Les résultats sont les suivants : Lot 1 $\Sigma p = 1875$ $\Sigma p^2 = 14325$ Lot 2 $\Sigma p = 2817$ $\Sigma p^2 = 18206$ Comparer les poids des adultes des deux lots. L'effet de la compétition larvaire est-il significatif ?

Ça serait bien plus simple si on avait les données. Parce que là, il faut savoir les formules. Mais est-ce bien important de savoir les formules ? Dans R ce n'est pas franchement essentiel, mais il faut avoir compris quelques fondements pour s'en servir. Conclusion : avec R on peut faire presque tous les anciens exercices et beaucoup de nouveaux !