# STA108 – Sondage aléatoire simple

*Cours n°2 du 02/10/20*

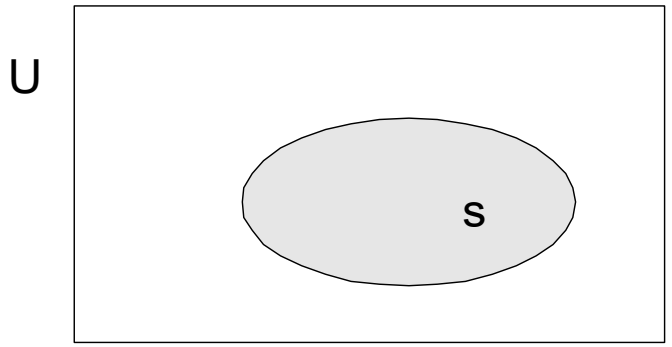
Sylvie Rousseau

le cnam

# Sommaire

# Définition

Tirage d'un échantillon *s* de *n* unités sans remise et à probabilités égales dans une population *U* de taille *N*



# Utilisations

- Mise en œuvre généralement aisée
- Ne nécessite pas d'information auxiliaire
- Référence, étalon auquel comparer les autres plans
- Brique de base pour des sondages plus complexes

# Exemple : estimer la surface moyenne d'exploitations agricoles

Soit une population de N = 5 exploitations agricoles de surfaces respectives : $y_1 = 10$, $y_2 = 30$, $y_3 = 40$, $y_4 = 60$ et $y_5 = 110$ ha.

$$\overline{Y} = 50 \qquad S_y^2 = 1450$$

On sélectionne 2 fermes selon un plan simple simple sans remise.

Il y a $C_5^2 = 10$ échantillons possibles :

*S ={ {1,2} ;{1,3} ;{1,4} ;{1,5} ;{2,3} ;{2,4} ;{2,5} ;{3,4} ;{3,5} ;{4,5} }*

Tous sont équiprobables : $\forall s \in S, \quad p(s) = 1/C_5^2 = 1/10$

Toutes les exploitations ont la même probabilité d'être sélectionnées :

$$P(k \in s) = \frac{4}{10} = \frac{2}{5} = \frac{n}{N}$$

# Sommaire


1) Généralités
**2) Notations**
3) Estimations de paramètres
4) Précision des estimateurs
5) Comment construire des intervalles de confiance ?
6) Comment déterminer la taille de l'échantillon ?
7) Formulaire
8) Application dans R et dans SAS

# Notations dans la population

- Population finie $U$ de $N$ objets identifiables (ou individus, unités statistiques) : $U = \{1,2,...,k,...,N\}$
- Variable d'intérêt $Y$ de caractéristique individuelle $Y_k$
- Total : $T_Y = \sum_{k \in U} Y_k$
- Moyenne : $\bar{Y} = \frac{T_Y}{N} = \frac{1}{N} \sum_{k \in U} Y_k$
- Variance : $\sigma_y^2 = \frac{1}{N} \sum_{k \in U} \left(Y_k - \bar{Y}\right)^2$
- Dispersion (variance modifiée) :

$$S_y^2 = \frac{1}{N-1} \sum_{k \in U} \left(Y_k - \bar{Y}\right)^2 = \frac{N}{N-1} \sigma_y^2$$

# Notations dans l'échantillon

- Échantillon *s* : sous-ensemble de *U*
- Ensemble des échantillons possibles : $\mathcal{S}$
- Plan de sondage probabiliste : loi de probabilité sur $\mathcal{S}$

$$p(s) \geq 0, \forall s \in S, \text{ et } \sum_{s \in S} p(s) = 1.$$

- Ici : $p(s) = 1/C_N^n$
- Moyenne empirique : $\hat{\bar{y}} = \frac{1}{n} \sum_{k \in S} Y_k$
- Dispersion empirique : $\hat{s}_y^2 = \frac{1}{n-1} \sum_{k \in S} \left( Y_k - \hat{\bar{y}} \right)^2$

# Notations dans l'échantillon

- Probabilité d'inclusion d'ordre un : $\pi_k = P(k \in s) = \sum_{s \in S / k \in s} p(s) = E(I_k)$

avec $I_k = \begin{cases} 1 & k \in S \\ 0 & k \notin S \end{cases}$ Variable indicatrice de Cornfield

Ici : $\forall k \in U,\ \pi_k = P(k \in s) = \frac{n}{N} = f$ = taux de sondage

*Remarque* : $\sum_{k \in U} \pi_k = n$ *plan de taille fixe*

- Probabilité d'inclusion d'ordre deux ou double de *k* et *l* ($k \neq l$) :

$$\pi_{kl} = p(k \in s, l \in s) = \sum_{s \in S / k, l \in s} p(s) = E(I_k I_l)$$

Ici : $\pi_{kl} = \frac{n}{N} \frac{n-1}{N-1}$

# Sommaire


1) Généralités
2) Notations
**3) Estimations de paramètres**
4) Précision des estimateurs
5) Comment construire des intervalles de confiance ?
6) Comment déterminer la taille de l'échantillon ?
7) Formulaire
8) Application dans R et dans SAS

# Comment estimer une moyenne ?

- Estimateur de la moyenne :

$$\hat{\bar{y}} = \frac{1}{n}\sum_{k \in S} Y_k = \hat{\bar{y}}(s)$$

- Exemple :

| *Echantillon s* | {1,2} | {1,3} | {1,4} | {1,5} | {2,3} | {2,4} | {2,5} | {3,4} | {3,5} | {4,5} |
|---|---|---|---|---|---|---|---|---|---|---|
| Moyenne $\hat{\bar{y}}$ | 20 | 25 | 35 | 60 | 35 | 45 | 70 | 50 | 75 | 85 |

- Propriété : estimateur sans biais $E\left(\hat{\bar{y}}\right) = \bar{Y}$

$$AN : E\left(\hat{\bar{y}}\right) = \sum_s p(s)\hat{\bar{y}}(s) = \frac{1}{10}\sum_s \hat{\bar{y}}(s) = 50\,ha = \bar{Y}$$

# Comment estimer un total ?

- Estimateur du total :

$$\hat{t}_y = N\,\hat{\bar{y}} = \frac{N}{n}\sum_{k\in S} Y_k = \hat{t}_y(s)$$

- Exemple :

| *Echantillon s* | {1,2} | {1,3} | {1,4} | {1,5} | {2,3} | {2,4} | {2,5} | {3,4} | {3,5} | {4,5} |
|---|---|---|---|---|---|---|---|---|---|---|
| *Total* $\hat{t}_y$ | 100 | 125 | 175 | 300 | 175 | 225 | 350 | 250 | 375 | 425 |

- Sans biais : $E\left(\hat{t}_y\right) = T_y$

$$AN : E\left(\hat{t}_y\right) = \sum_{s\in S} p(s)\,\hat{t}_y(s) = \frac{2500}{10} = 250\,ha = T_y$$

# Comment estimer une proportion ?

***Cas particulier d'une moyenne***

- Vraie proportion (inconnue) à estimer $p = N_0 / N$
- Estimateur :

$$\hat{p} = \frac{1}{n} \sum_{k \in s} y_k = \frac{n_0}{n}$$

- Exemple : proportion de fermes de plus de 50 ha : *p* = 40%

| *Echantillon s* | {1,2} | {1,3} | {1,4} | {1,5} | {2,3} | {2,4} | {2,5} | {3,4} | {3,5} | {4,5} |
|---|---|---|---|---|---|---|---|---|---|---|
| *Proportion* $\hat{p}$ | 0 | 0 | 0,5 | 0,5 | 0 | 0,5 | 0,5 | 0,5 | 0,5 | 1 |

- Sans biais : $E(\hat{p}) = p$

$$AN : E(\hat{p}) = \sum_{s \in S} p(s)\,\hat{p}(s) = 0{,}4 = p$$

# Sommaire


1) Généralités
2) Notations
3) Estimations de paramètres
**4) Précision des estimateurs**
5) Comment construire des intervalles de confiance ?
6) Comment déterminer la taille de l'échantillon ?
7) Formulaire
8) Application dans R et dans SAS

# Précision de l'estimateur d'une moyenne

- Des fluctuations dues à l'échantillonnage
- Précision de l'estimateur d'une moyenne :

$$Var\left(\hat{\bar{y}}\right) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$$

- Exemple :

| *Echantillon s* | *{1,2}* | *{1,3}* | *{1,4}* | *{1,5}* | *{2,3}* | *{2,4}* | *{2,5}* | *{3,4}* | *{3,5}* | *{4,5}* |
|---|---|---|---|---|---|---|---|---|---|---|
| Moyenne $\hat{\bar{y}}$ | 20 | 25 | 35 | 60 | 35 | 45 | 70 | 50 | 75 | 85 |

$$AN: \quad Var\left(\hat{\bar{y}}\right) = \sum_{s \in S} p(s)\left[\hat{\bar{y}}(s) - \bar{Y}\right]^2 = 435 \simeq (20{,}9\,ha)^2$$

$$= (1 - 2/5) * (1450/2) = 435$$

# Précision de l'estimateur d'un total ou d'une proportion

- Précision de l'estimateur d'un total :

$$Var\left(\hat{t}_y\right) = N^2\left(1-\frac{n}{N}\right)\frac{S_y^2}{n}$$

- Précision de l'estimateur d'une proportion :

$$Var\left(\hat{p}\right) = \left(1-\frac{n}{N}\right)\frac{N}{N-1}\frac{p(1-p)}{n}$$

# La variance d'échantillonnage dépend de 3 facteurs

- La taille de l'échantillon

  Plus l'échantillon est grand, plus la précision est importante

- La dispersion de la variable d'intérêt dans la population

  Plus la population est hétérogène, plus les fluctuations d'échantillonnage sont élevées

- Le taux de sondage

  Plus il est élevé, plus on limite l'aléa d'échantillonnage

# Estimer la précision de l'estimateur d'une moyenne

- Vraie variance d'échantillonnage (inconnue) de l'estimateur d'une moyenne :

$$Var\left(\hat{\bar{y}}\right) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$$

- Estimée sans biais (n >1) par : $\hat{Var}\left(\hat{\bar{y}}\right) = \left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$
- Exemple :

| s | {1,2} | {1,3} | {1,4} | {1,5} | {2,3} | {2,4} | {2,5} | {3,4} | {3,5} | {4,5} | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\bar{y}}$ | 20 | 25 | 35 | 60 | 35 | 45 | 70 | 50 | 75 | 85 | 500 |
| $\hat{s}_y^2$ | 200 | 450 | 1250 | 5000 | 50 | 450 | 3200 | 200 | 2450 | 1250 | 14500 |
| $\hat{Var}\left(\hat{\bar{y}}\right)$ | 60 | 135 | 375 | 1500 | 15 | 135 | 960 | 60 | 735 | 375 | 4350 |

Et on vérifie : $E\left[\hat{Var}\left(\hat{\bar{y}}\right)\right] = 435 = Var\left(\hat{\bar{y}}\right)$

le cnam

# Estimer la précision de l'estimateur d'un total ou d'une proportion

- Estimateur sans biais (n >1) de la variance d'échantillonnage de l'estimateur d'un total :

$$\hat{V}ar\left(\hat{t}_y\right) = N^2\left(1-\frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$$

- Estimateur sans biais (n >1) de la variance d'échantillonnage de l'estimateur d'une proportion :

$$\hat{V}ar(\hat{p}) = \left(1-\frac{n}{N}\right)\frac{\hat{p}\left(1-\hat{p}\right)}{n-1}$$

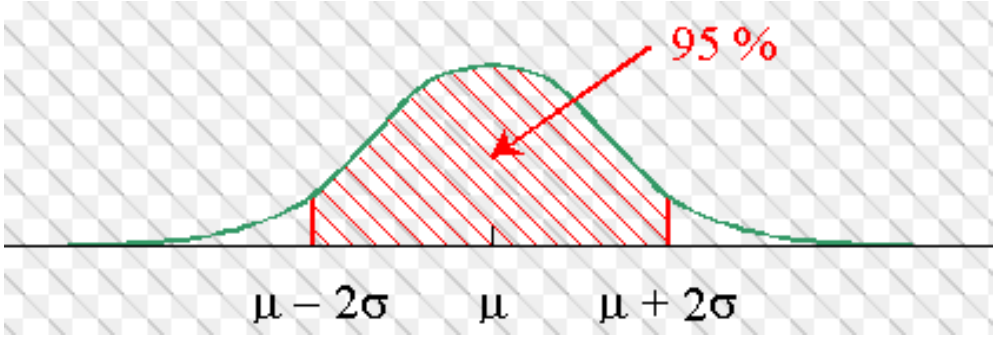
# Sommaire


1) Généralités
2) Notations
3) Estimations de paramètres
4) Précision des estimateurs
**5) Comment construire des intervalles de confiance ?**
6) Comment déterminer la taille de l'échantillon ?
7) Formulaire
8) Application dans R et dans SAS

# Intervalles de confiance

Un intervalle de confiance à 95% est une fourchette autour d'une valeur estimée qui a 95 chances sur 100 d'intercepter la vraie valeur.

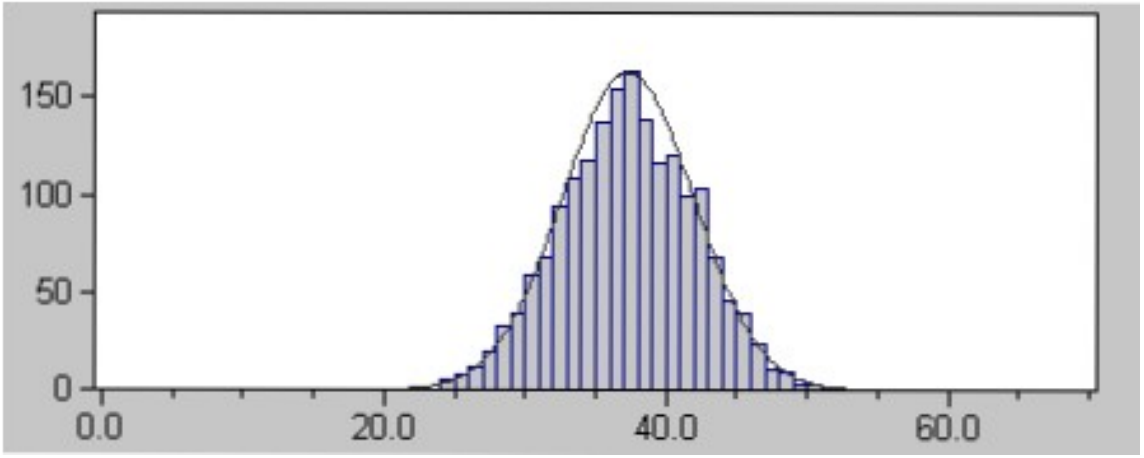
*C'est à dire que si l 'on faisait 100 estimations avec la même méthode de tirage et sur des échantillons de même taille,on en aurait 95/100 qui intercepteraient la vraie valeur.*




- En général intervalle de confiance symétrique autour d'une valeur centrale
- Nécessite de connaître au moins approximativement la distribution de probabilité de l'estimateur
- La longueur de l'intervalle diminue avec n et augmente avec la variance de l'estimateur et avec le niveau de confiance

# Le théorème central limite

- La moyenne d'un échantillon de n observations indépendantes issues d'une population de moyenne $\mu$ et d'écart-type $\sigma$ converge, si n augmente, vers une loi normale $N(\mu, \sigma^2/n)$
- Illustration animée : http://www.vias.org/simulations/simusoft_cenlimit.html
- n>30 est souvent suffisant



# Comment construire des intervalles de confiance ?

- Hypothèse (*n* grand) : $$\frac{\hat{\bar{y}} - \bar{Y}}{\sqrt{Var\,(\hat{\bar{y}})}} \rightarrow N(0,\,1)$$
- Intervalle au niveau de confiance 95% pour la moyenne :

$$IC_{95\%}\left(\bar{Y}\right) = \left[\hat{\bar{y}} - 1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)},\ \ \hat{\bar{y}} + 1{,}96\sqrt{\hat{Var}\left(\hat{\bar{y}}\right)}\right]$$

- A.N. : pour *s= {1,2}* , $IC_{95\%}\left(\bar{Y}\right) = \left[20 \pm 15\right]$ *ha*
- Même principe pour un total ou une proportion

# Sommaire


1) Généralités
2) Notations
3) Estimations de paramètres
4) Précision des estimateurs
5) Comment construire des intervalles de confiance ?
**6) Comment déterminer la taille de l'échantillon ?**
7) Formulaire
8) Application dans R et dans SAS

# Comment déterminer la taille d'un échantillon ?

- La taille d'un échantillon est déterminée par :
  - le budget disponible

    Si budget disponible très serré, alors *n* = budget disponible divisé par le coût unitaire
  - la précision souhaitée

    Choix de *n* assurant une précision minimale

# Taille minimale d'échantillon pour une précision donnée

- pour une **erreur absolue *b*** acceptée : $1{,}96^2\left(\frac{1}{n}-\frac{1}{N}\right)S_Y^2 \le b^2$

$$n_{\min} \ge \frac{1}{\frac{1}{N}+\frac{b^2}{1{,}96^2 S_y^2}}$$

Si *n/N* est négligeable : $n_{\min} \ge \frac{1{,}96^2 S_y^2}{b^2}$

- pour une **erreur relative *β*** acceptée $1{,}96^2\left(\frac{1}{n}-\frac{1}{N}\right)\frac{S_Y^2}{\overline{Y}^2} \le \beta^2$

$$n_{\min} \ge \frac{1}{\frac{1}{N}+\frac{\beta^2}{1{,}96^2 CV_y^2}}$$

Si *n/N* est négligeable : $n_{\min} \ge \frac{1{,}96^2 CV_y^2}{\beta^2}$

# Taille minimale d'échantillon pour une précision donnée

- La taille de l'échantillon doit être un entier
- $S_y^2$ ou $CV_y$ sont inconnues …

Les estimer : comment ?

- par une enquête précédente sur le même thème
- par une enquête portant sur une variable x liée à y
- par une enquête préalable de petite taille
- par avis d'expert
- par référence à une proportion et en envisageant la pire des situations (p=0,5)

# Erreur absolue pour une proportion selon la taille de l'échantillon

- Précision d'une enquête selon *n* et *p* (*N* grand)

| n \ p | 0,1 | 0,2 | 0,3 | 0,4 | 0,5 |
|---|---|---|---|---|---|
| **10** | 0,186 | 0,248 | 0,284 | 0,304 | 0,310 |
| **100** | 0,059 | 0,078 | 0,090 | 0,096 | 0,098 |
| **500** | 0,026 | 0,035 | 0,040 | 0,043 | 0,044 |
| **1 000** | 0,019 | 0,025 | 0,028 | 0,030 | 0,031 |
| **2 000** | 0,013 | 0,018 | 0,020 | 0,021 | 0,022 |
| **5 000** | 0,008 | 0,011 | 0,013 | 0,014 | 0,014 |
| **10 000** | 0,006 | 0,008 | 0,009 | 0,010 | 0,010 |

$$1.96 \times \sqrt{\frac{p(1-p)}{n}}$$

*Note de lecture :*

*Dans un échantillon de taille 1 000, une proportion de 0,5 est connue à ± 0,03 près au niveau de confiance 95%*

# Sommaire


1) Généralités
2) Notations
3) Estimations de paramètres
4) Précision des estimateurs
5) Comment construire des intervalles de confiance ?
6) Comment déterminer la taille de l'échantillon ?
**7) Formulaire**
8) Application dans R et dans SAS

# Formulaire

$$p(s) = 1 / C_N^n$$

$$\forall k \in U,\ \pi_k = P(k \in s) = \frac{n}{N} = f$$

| Paramètre d'intérêt / Statistique | Moyenne | Proportion $p = N_0/N$ | Total |
|---|---|---|---|
| Estimateur du paramètre d'intérêt | $\hat{\bar{y}} = \frac{1}{n}\sum_{k \in S} Y_k = \hat{\bar{y}}(s)$ | $\hat{p} = \frac{1}{n}\sum_{k \in s} y_k = \frac{n_0}{n}$ | $\hat{t}y = N \times \hat{\bar{y}} = \frac{N}{n}\sum_{k \in S} Y_k$ |
| Vraie variance d'échantillonnage de cet estimateur | $Var(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$ | $Var(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{N}{N-1}\frac{p(1-p)}{n}$ | $Var(\hat{t}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{S_y^2}{n}$ |
| Estimateur de la variance d'échantillonnage | $\hat{Var}(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$ | $\hat{Var}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}$ | $\hat{Var}(\hat{t}_y) = N^2\left(1 - \frac{n}{N}\right)\frac{\hat{s}_y^2}{n}$ |

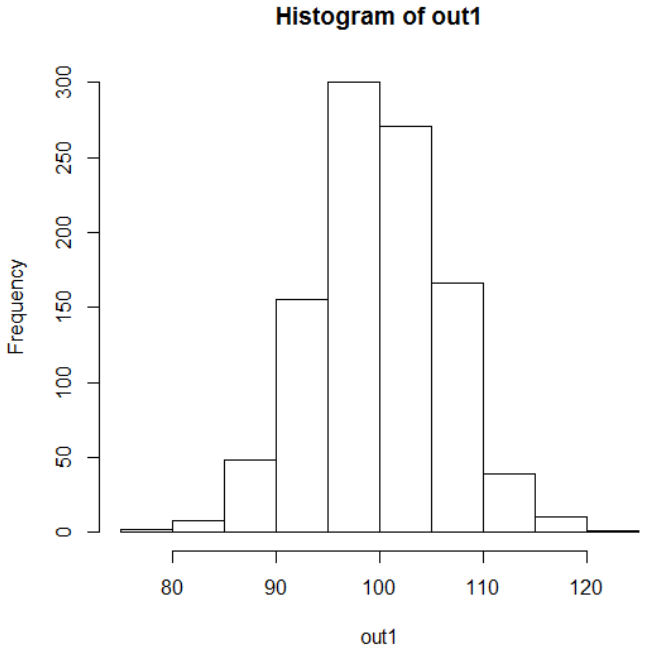
# Sommaire


1) Généralités
2) Notations
3) Estimations de paramètres
4) Précision des estimateurs
5) Comment construire des intervalles de confiance ?
6) Comment déterminer la taille de l'échantillon ?
7) Formulaire
**8) Application dans R et dans SAS**

# Application dans R

- On crée une population de 1001 valeurs entre 50 et 150 par intervalles de 0.1
- La moyenne est 100
- On crée 1000 échantillons de taille 20 et on calcule la moyenne sur chacun
- La moyenne des moyennes est proche de 100

```
> x <- seq(50,150,by=0.1)
> mean(x)
[1] 100
>
> # 1000 echantillons de taille 20
> out1 <- replicate(1000, mean(sample(x,20)))
> mean(out1)
[1] 99.86042
> head(out1)
[1] 102.965 100.835  95.800 100.680  89.365  90.520
> var(out1)
[1] 40.32327
> hist(out1)
```




le cnam

# Application dans R

- Le script R suivant crée un vecteur avec les données, puis calcule toutes les combinaisons de taille 2, donne les moyennes et la moyenne générale

```
> x <- c(5,8,10,12,15)
> mean(x)
[1] 10
> samples <- combn(x,2)
> samples
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    5    5    5    5    8    8    8   10   10    12
[2,]    8   10   12   15   10   12   15   12   15    15
> fix(xbars)
> head(xbars)
[1]  6.5  7.5  8.5 10.0  9.0 10.0
> xbars <- colMeans(samples)
> mean(xbars)
[1] 10
```

# Application dans R

```
> # la fonction sample sur R
> # sample(x, size, replace = FALSE, prob = NULL)
> y <- 1:20
> head(y)
[1] 1 2 3 4 5 6
> sample(y,10)
 [1] 19  8  1  2 13 11 17 16  6  4
> sample(y,10, replace=TRUE)
 [1]  1 11 13 14 20 18  3  1  4  6
>
> # fixer les générateurs de nombres aléatoires
> set.seed(123456)
> sample(y,10)
 [1] 16 15  8  6 17  3 18  2 12 13
> set.seed(123456)
> sample(y,10)
 [1] 16 15  8  6 17  3 18  2 12 13
>
> # théorème central limite
> u <- runif(1000)
> head(u)
[1] 0.7979891 0.5937940 0.9053100 0.8808486 0.9938366 0.8959563
> out3 <- replicate(1000, mean(sample(u,20)))
> hist(out3)
>
> out4 <- replicate(10000, mean(sample(u,50)))
> hist(out4)
```

R intègre plusieurs fonctions pour le tirage et l'estimation, par exemple *sample*

Il y a aussi plusieurs packages » : http://cran.r-project.org/web/views/OfficialStatistics.html

- « sampling » de Yves Tillé et Alina Matéi : http://cran.r-project.org/web/packages/sampling/index.html
- « survey » de Thomas Lumley : http://cran.r-project.org/web/packages/survey/index.html

# Application dans SAS

## Tirage d'un échantillon

```
PROC  SURVEYSELECT
        DATA = base de sondage
        METHOD = SRS
        {SAMPSIZE=taille de l'échantillon
         SAMPRATE=taux de sondage}
        OUT = table échantillon
        STATS
        OUTSIZE  ;
    ID liste de variables ;
RUN;
```

# Application dans SAS

## Estimation des paramètres

**PROC SURVEYMEANS**

DATA = *table échantillon*

RATE=*taux de sondage*
TOTAL=*taille de la population*

MISSING

*statistiques* ;

WEIGHT *variable de pondération* ;

VAR *liste de variables à estimer* ;

CLASS *liste de variables catégorielles à estimer* ;

RUN;