MAT 3777
Échantillonnage et sondages

Chapitre 2
Échantillonnage aléatoire simple

P. Boily (uOttawa)

Session d'hiver – 2022

Aperçu

2.1 – Motivation

Soit $\mathcal{U}$ une population composée de N unités, dont les réponses resp. sont

$$\mathcal{U} = \{u_1, \ldots, u_N\}.$$

Supposons que l'on s'intéresse à la **moyenne** μ d'une population cible $\mathcal{U}$, où

$$\mu = \frac{1}{N} \sum_{j=1}^{N} u_j.$$

Puisque la population est de taille finie, il est possible d'évaluer μ directement... .

En pratique, nous n'avons que rarement accès aux réponses pour la population $\mathcal{U}$ dans son entiereté, d'où le recours aux

.

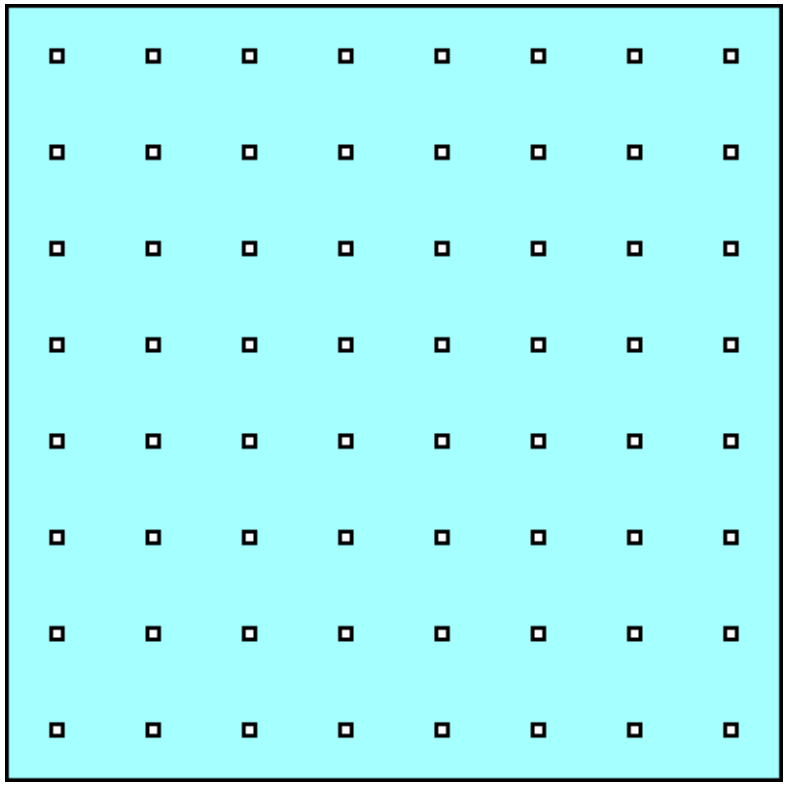
Un **échantillon** $\mathcal{Y}$ de taille $n \leq N$ est un sous-ensemble de la population cible $\mathcal{U}$,

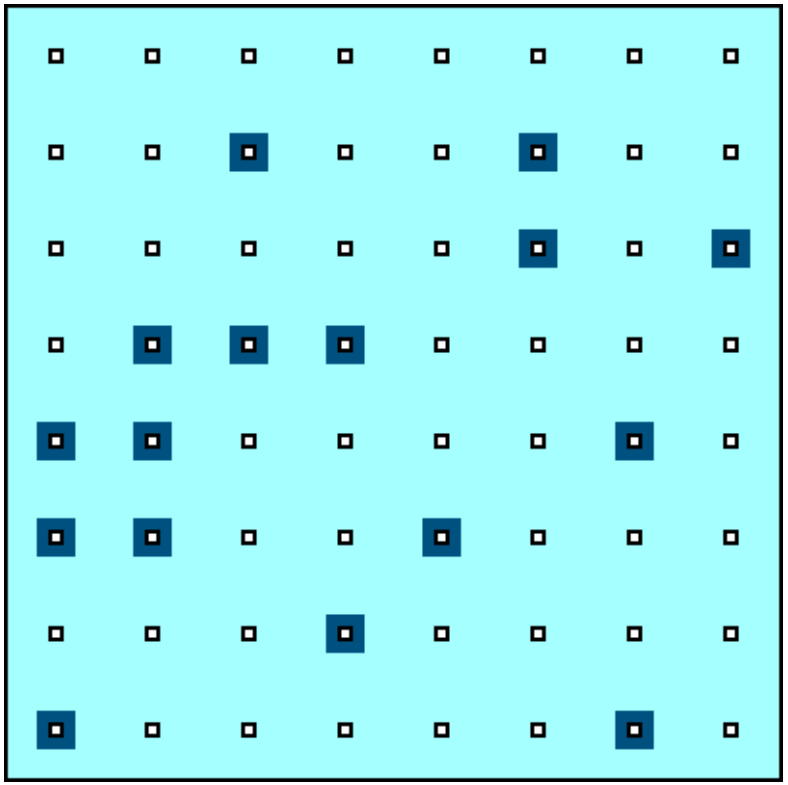
$$\mathcal{Y} \subseteq \{y_1, \ldots, y_n\} \subseteq \{u_1, \ldots, u_N\} = \mathcal{U},$$

à partir duquel on peut approcher μ à l'aide de la **moyenne d'échantillon**

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

(ce n'est pas la seule méthode).





On obtient un (EAS) de taille n en prélevant n unités de la population cible, .

À chaque stade de l'échantillonnage,

.

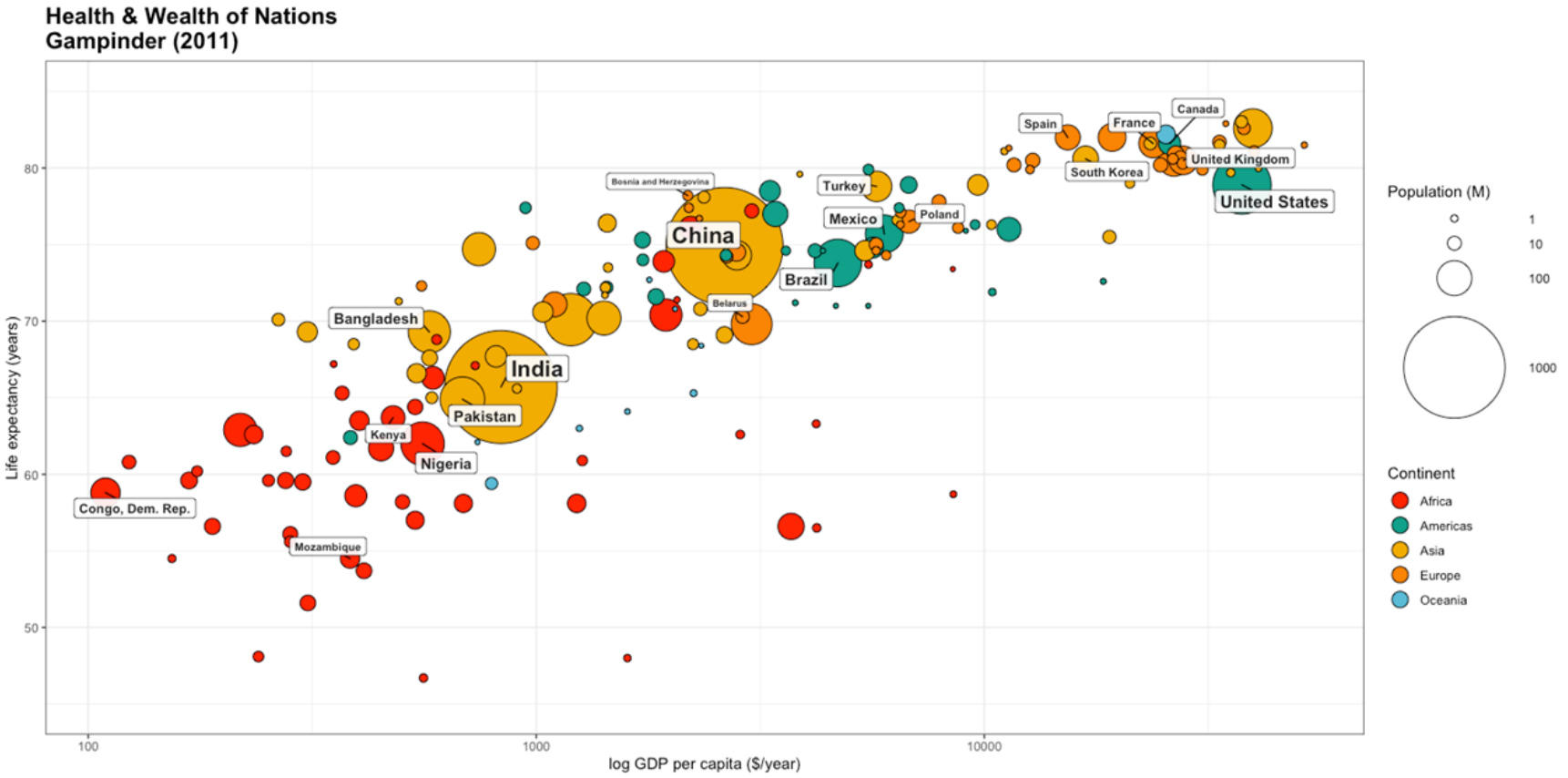
Dans un plan d'échantillonnage EAS, chaque sous-ensemble de taille n .

Comment choisir un échantillon **aléatoire**? À l'époque, on les choisissait "à la mitaine", à l'aide de tables de nombres aléatoires. De nos jours, on se sert de logiciels (SAS, R, etc.) afin de choisir un échantillon aléatoire.

Exemple: Quelle est la durée de vie moyenne par pays en 2011?

```
# preparation de l'ensemble de donnees
> setwd("C:/Users/idlew/Documents/Courses/ES")
> gapminder = read.csv("DATA/gapminder.csv")
> str(gapminder)
```

```
'data.frame': 10545 obs. of 9 variables:
 $ country : chr "Albania" "Algeria" "Angola" ...
 $ year : int 196 196 196 196 196 196 196 196 ...
 $ infant_mortality: num 115.4 148.2 208 NA 59.9 ...
 $ life_expectancy : num 62.9 47.5 36 63 65.4 ...
 $ fertility : num 6.19 7.65 7.32 4.43 3.11 4.55 4.82 3.45 ...
 $ population : int 1636054 11124892 5270844 54681 20619075 ...
 $ gdp : num NA 1.38e+1 NA NA 1.08e+11 ...
 $ continent : chr "Europe" "Africa" "Africa" "Americas" ...
 $ region : chr "Southern Europe" "Northern Africa" ...
```

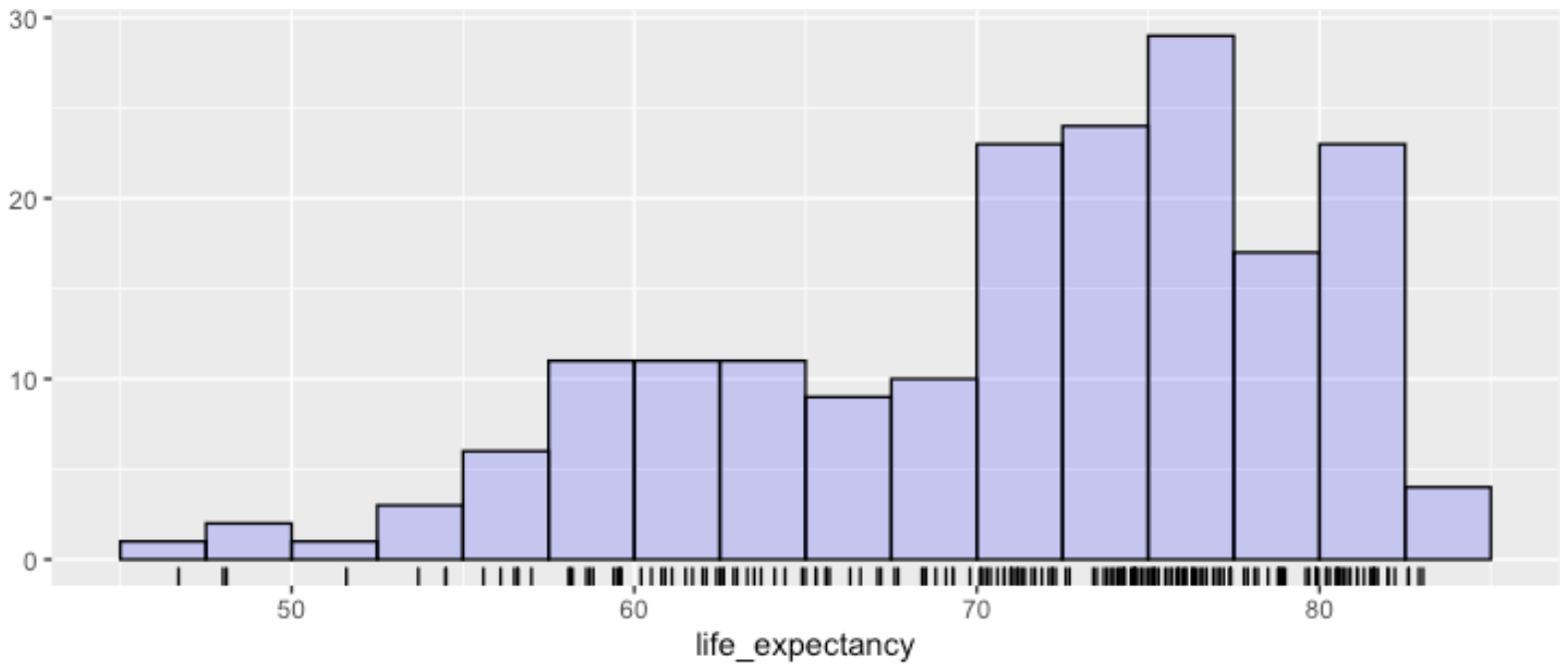

Health & Wealth of Nations
Gampinder (2011)
Life expectancy (years)
80
70
60
50
100
1000
10000
log GDP per capita ($/year)
Spain
France
Canada
United Kingdom
South Korea
United States
Bosnia and Herzegovina
Turkey
Mexico
Poland
China
Brazil
Belarus
Bangladesh
India
Pakistan
Kenya
Nigeria
Congo, Dem. Rep.
Mozambique
Population (M)
1
10
100
1000
Continent
Africa
Americas
Asia
Europe
Oceania

```
# extraction des donnees d'interet: life_expectancy, year=2011
> library(tidyverse)
> gapminder.EAS <- gapminder %>%
 filter(year==2011) %>%
 select(life_expectancy)
> summary(gapminder.EAS)
> gapminder.EAS[,]
```

```
 life_expectancy
 Min. :46.7
 1st Qu.:65.3
 Median :73.7
 Mean :71.18
 3rd Qu.:77.4
 Max. :83.0
```

```
  [1] 77.4 76.1 58.1 75.9 76.0 73.5 75.2 82.2 80.7 70.8 72.6 78.8 69.3
 [14] 75.2 70.3 80.2 71.2 61.1 71.7 72.1 78.2 56.5 73.8 76.9 74.2 59.5
 [27] 60.8 67.6 58.1 81.6 71.4 48.1 56.1 78.9 74.9 77.0 67.2 58.8 60.9
 [40] 79.9 57.0 77.1 77.9 81.1 77.8 79.9 62.5 74.6 75.3 70.4 74.3 58.7
 [53] 60.2 76.3 62.9 65.3 80.3 81.6 75.8 63.3 67.1 72.2 80.5 63.5 80.5
 [66] 71.2 71.0 71.6 58.2 54.5 65.6 62.4 72.2 83.0 75.0 82.9 65.7 70.1
 [79] 74.1 67.7 80.6 81.6 82.0 74.6 82.6 78.1 69.1 63.7 62.1 80.6 79.0
 [92] 68.5 65.0 74.6 76.6 46.7 61.5 60.5 74.3 81.5 80.0 75.6 62.6 56.6
[105] 74.6 79.6 59.6 81.3 68.8 73.7 75.7 68.4 72.3 65.6 76.7 73.9 54.5
[118] 62.6 69.3 80.9 75.9 80.8 77.4 59.6 62.0 81.1 76.3 64.9 77.4 59.4
[131] 74.0 78.5 70.2 76.5 80.2 77.4 79.7 74.5 69.8 65.3 74.6 71.0 72.7
[144] 78.9 64.4 75.1 73.4 55.6 81.5 76.1 79.9 63.0 56.6 82.0 76.4 66.3
[157] 71.0 48.0 81.7 82.6 75.1 70.1 61.7 74.3 71.3 59.6 70.8 71.9 77.2
[170] 78.8 68.5 58.6 71.1 75.5 80.5 78.9 76.3 70.6 64.1 74.8 74.2 74.7
[183] 66.6 53.7 51.6
```

```
> ggplot(data=gapminder.EAS, aes(life_expectancy)) + geom_rug() +
    geom_histogram(col="black", fill="blue", alpha=.2,
                   breaks=seq(45, 85, by = 2.5))
```



On voit qu'en 2011, la moyenne pour les $N =$ pays est $\mu =$.

On prélève au hasard un échantillon de taille $n = 10$, par exemple:

```
# replicabilite
> set.seed(1234) # sur un PC
> N = dim(gapminder.EAS)[1]
> n = 10
> (sample.ind = sample(1:N,n, replace=FALSE))
```

```
 [1] 28 80 150 101 111 137 133 166 144 132
```

```
> (gapminder.EAS.n = gapminder.EAS[sample.ind,])
```

```
 [1] 67.6 67.7 76.1 80.0 75.7 79.7 70.2 59.6 78.9 78.5
```

```
  [1] 77.4 76.1 58.1 75.9 76.0 73.5 75.2 82.2 80.7 70.8 72.6 78.8 69.3
 [14] 75.2 70.3 80.2 71.2 61.1 71.7 72.1 78.2 56.5 73.8 76.9 74.2 59.5
 [27] 60.8 67.6 58.1 81.6 71.4 48.1 56.1 78.9 74.9 77.0 67.2 58.8 60.9
 [40] 79.9 57.0 77.1 77.9 81.1 77.8 79.9 62.5 74.6 75.3 70.4 74.3 58.7
 [53] 60.2 76.3 62.9 65.3 80.3 81.6 75.8 63.3 67.1 72.2 80.5 63.5 80.5
 [66] 71.2 71.0 71.6 58.2 54.5 65.6 62.4 72.2 83.0 75.0 82.9 65.7 70.1
 [79] 74.1 67.7 80.6 81.6 82.0 74.6 82.6 78.1 69.1 63.7 62.1 80.6 79.0
 [92] 68.5 65.0 74.6 76.6 46.7 61.5 60.5 74.3 81.5 80.0 75.6 62.6 56.6
[105] 74.6 79.6 59.6 81.3 68.8 73.7 75.7 68.4 72.3 65.6 76.7 73.9 54.5
[118] 62.6 69.3 80.9 75.9 80.8 77.4 59.6 62.0 81.1 76.3 64.9 77.4 59.4
[131] 74.0 78.5 70.2 76.5 80.2 77.4 79.7 74.5 69.8 65.3 74.6 71.0 72.7
[144] 78.9 64.4 75.1 73.4 55.6 81.5 76.1 79.9 63.0 56.6 82.0 76.4 66.3
[157] 71.0 48.0 81.7 82.6 75.1 70.1 61.7 74.3 71.3 59.6 70.8 71.9 77.2
[170] 78.8 68.5 58.6 71.1 75.5 80.5 78.9 76.3 70.6 64.1 74.8 74.2 74.7
[183] 66.6 53.7 51.6
```

La moyenne (empirique) de cet échantillon est

$$\overline{y} = \qquad .$$

Un échantillon différent peut mener à une estimation différente:

```
> set.seed(12345) # sur un mac
> (sample.ind = sample(1:N,n, replace=FALSE))
```

```
 [1] 134 162 140 184 83 30 59 91 129 175
```

```
> (gapminder.EAS.n = gapminder.EAS[sample.ind,])
```

```
 [1] 76.5 70.1 65.3 53.7 82.0 81.6 75.8 79.0 77.4 80.5
```

La moyenne (empirique) de ce second échantillon est

```
> mean(gapminder.EAS.n)
```

```
[1] 74.194
```

C'est tout à fait normal –

.

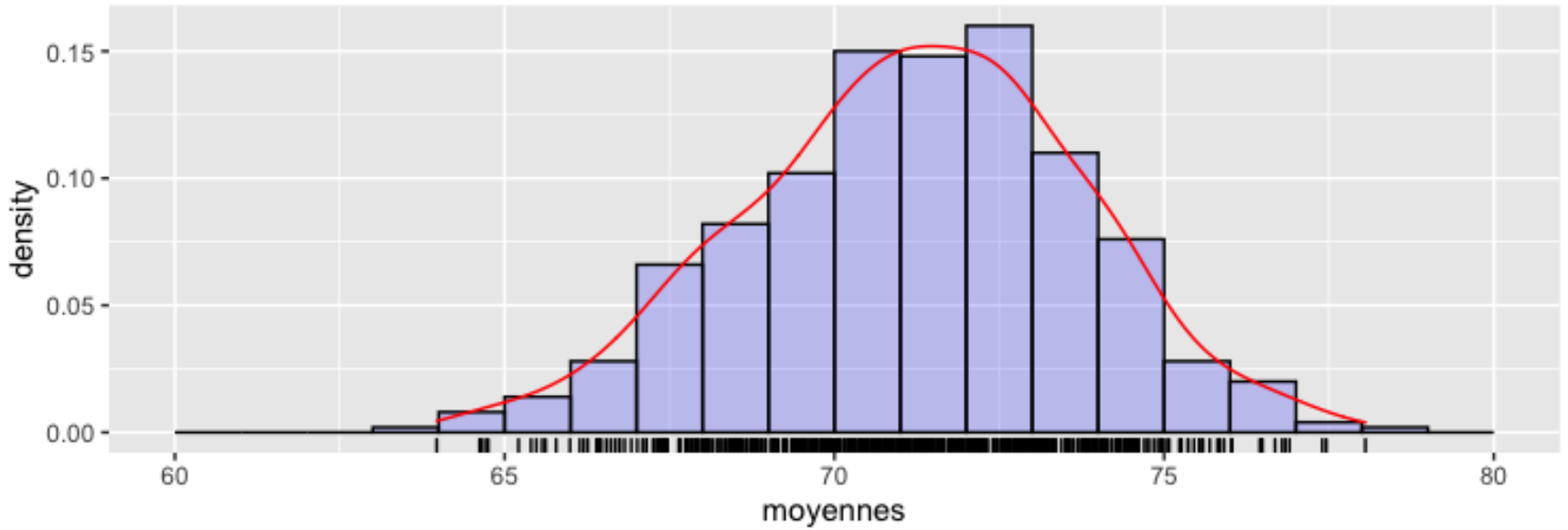
La **variabilité d'échantillonnage** explique comment les estimations varient en fonction de l'échantillon.

Par exemple, si on prépare $m = 500$ échantillons, chacun de taille $n = 10$, on pourrait obtenir les moyennes empiriques de la page suivante:

```
> set.seed(12) # sur un mac
> N=dim(gapminder.EAS)[1]
> n=10
> m=500
> moyennes <- c()
> for(k in 1:m){
 moyennes[k] <- mean(gapminder.EAS[sample(1:N,n, replace=FALSE),])
 }
> moyennes
```

```
 [1] 72.219 68.110 68.770 73.260 67.430 72.864 71.900 70.054 72.477 69.930
[11] 73.100 70.300 68.957 71.060 69.250 74.840 64.700 74.942 68.840 73.780
[21] 69.390 68.672 67.250 72.080 72.650 69.540 70.040 69.720 72.080 72.140
...
```

```
> ggplot(data=data.frame(moyennes), aes(moyennes)) +
    geom_histogram(aes(y =..density..), breaks=seq(60, 80, by = 1),
      col="black", fill="blue", alpha=.2) +
    geom_density(col=2) + geom_rug(aes(moyennes))
```



2.2 – Concepts fondamentaux

La **variance de la population** σ^2 est une mesure de **dispersion**, c-à-d de la tendance qu'ont les valeurs de la réponse à s'éloigner de la **moyenne de la population** μ:

$$\sigma^2 = \frac{1}{} \sum_{j=1}^{N} (u_j - \mu)^2 = \frac{1}{} \sum_{j=1}^{N} (u_j^2 - 2u_j\mu + \mu^2)$$

$$= \frac{1}{} \left(\sum_{j=1}^{N} u_j^2 - 2\mu \sum_{j=1}^{N} u_j + N\mu^2 \right) = \frac{1}{} \left(\sum_{j=1}^{N} u_j^2 - 2N\mu^2 + N\mu^2 \right)$$

$$= \frac{1}{} \sum_{j=1}^{N} \left(u_j^2 - N\mu^2 \right) = \frac{1}{} \sum_{j=1}^{N} u_j^2 - \mu^2$$

Les paramètres μ et σ^2 peuvent être interprétés en termes de l'
et de la d'une variable aléatoire.

Soit X une variable aléatoire discrète dont la est $f(x) = P(X = x)$. Ainsi,

$$\mathrm{E}[X] = \sum_x x f(x), \quad \mathrm{V}[X] = \sum_x (x - \mathrm{E}[X])^2 f(x), \quad \mathrm{SD}[X] = \sqrt{\mathrm{V}[X]}.$$

Pour un échantillon de taille $n = 1$ provenant de cette population, dont la valeur est représentée par la v.a. Y_1, nous avons $f(u_j) = P(Y_1 = u_j) =$, pour $j = 1, \ldots, N$, d'où

$$\mathrm{E}[Y_1] = \quad ,$$

et

$$\mathrm{V}[Y_1] = \qquad\qquad , \quad \mathrm{SD}[Y_1] = \qquad\qquad .$$

En général, l'estimateur $\overline{y}$ de la moyenne de population μ est produit à l'aide de – différents échantillons de taille n donne naissance à différents valeurs de $\overline{y}$.

Afin de contrôler l'erreur d'échantillonnage associé à un EAS, il faut connaître la ; en particulier, $\mathrm{E}[\overline{Y}]$ et $\mathrm{V}[\overline{Y}]$.

Si $y_1, \ldots, y_n$ sont des v.a.
, le **théorème de la limite centrée** impose $\overline{Y} \sim$.

Exemple: Considérons une population finie à $N = 4$ éléments:

$$u_1 = 2, \quad u_2 = 0, \quad u_3 = 1, \quad u_4 = 5.$$

La moyenne et la variance de population sont, resp.,

$\mu =$ et $\sigma^2 =$.

Supposons que l'on souhaite prélever de cette population un EAS sans remise de taille $n = 3$ afin d'estimer la moyenne μ.

Il y a tels échantillons.

Échantillon	**Valeurs**	$\overline{y}$	$P(\overline{Y} = \overline{y})$
u_1, u_2, u_3	$2, 0, 1$	1	$1/4$
u_1, u_2, u_4	$2, 0, 5$	$7/3$	$1/4$
u_1, u_3, u_4	$2, 1, 5$	$8/3$	$1/4$
u_2, u_3, u_4	$0, 1, 5$	2	$1/4$

Alors

$$\mathrm{E}[\overline{Y}] = \sum_{\overline{y}} \overline{y} P(\overline{Y} = \overline{y}) =$$

$$\mathrm{V}[\overline{Y}] = \sum_{\overline{y}} \overline{y}^2 P(\overline{Y} = \overline{y}) - \mathrm{E}^2[\overline{Y}] =$$

Mais on remarque que . Qu'est-ce qui se passe?

Voici comment on explique ce résultat.

Soit $\mathcal{U} = \{u_1, \ldots, u_N\}$ une population finie. On y prélève un EAS sans remise de taille n.

Soit Y_i la variable aléatoire qui représente la valeurs que prend la $i-$ème unité de l'EAS, respectivement.

Tous les Y_i ont des : pour tout $u_j \in \mathcal{U}$, nous obtenons (⚠ ne pas confondre l'unité u_j et sa réponse u_j):

$$P(Y_1 = u_j) =$$

$$P(Y_2 = u_j) =$$

$$P(Y_3 = u_j) =$$

$$=$$

et ainsi de suite: $P(Y_i = u_j) = \frac{1}{N}$ pour tout $1 \leq i \leq n$, $1 \leq j \leq N$, d'où $\mathrm{E}[Y_i] = \mu$ and $\mathrm{V}[Y_i] = \sigma^2$ pour tout i. Ainsi, à l'exemple précédent,

$$\mathrm{E}[Y_1] = \mathrm{E}[Y_2] = \mathrm{E}[Y_3] = \mu = 2 \quad \text{et} \quad \mathrm{V}[Y_1] = \mathrm{V}[Y_2] = \mathrm{V}[Y_3] = \sigma^2 = \frac{7}{2}.$$

Mais les $\{Y_i\}$ ne sont pas puisque (e.g.)

$$\mathrm{E}[\overline{Y}] = \mu = 2, \quad \text{mais } \mathrm{V}[\overline{Y}] = \mathrm{V}\left[\tfrac{Y_1+Y_2+Y_3}{3}\right] = \frac{7}{18} \neq \frac{\sigma^2}{3} = \frac{7/2}{3} = \frac{7}{6}.$$

C'est dans la variance que l'on observe une différence.

La **covariance** entre deux variables aléatoires (discrètes) $X_1,\ X_2$ est une .

Si $\mathrm{E}[X_i] = \mu_i$ et $\mathrm{V}[X_i] = \sigma_i^2 < \infty$, alors

$$\mathrm{Cov}[X_1, X_2] = .$$

Si X_1, X_2 peuvent toutes deux prendre des valeurs dans $\mathcal{U} = \{u_1, \ldots, u_N\}$, alors l'**espérance conjointe** est

$$\mathrm{E}[X_1 X_2] = .$$

Dans le cas où $X_1 = Y_i$ et $X_2 = Y_\ell$ (avec l'interprétation donnée en p. 22), pour $1 \leq i \neq \ell \leq n$, nous obtenons

$$P(Y_i = u_j, Y_\ell = u_k) =$$

$$=$$

Mais $\mathrm{E}[Y_i] = \mathrm{E}[Y_\ell] = \mu$, d'où

$$\mathrm{Cov}(Y_i, Y_\ell) = \qquad .$$

On se sert des propriétés $\sum u_\xi =$ et $\sum u_\xi =$ afin de simplifier l'expression quand $i \neq \ell$:

$$\mathrm{Cov}(Y_i, Y_\ell) =$$

$$=$$

$$=$$

$$= \qquad\qquad = -\frac{\sigma^2}{N-1}.$$

En utilisant les formules en p. 38 (chapitre 1), on obtient alors

$$\mathrm{E}[\overline{Y}] =$$

$$= \qquad\qquad = \mu, \quad \text{et}$$

$$\mathrm{V}[\overline{Y}] =$$

$$=$$

$$= \qquad\qquad = \frac{\sigma^2}{n}\Big(\frac{N-n}{N-1}\Big).$$

Retournons à notre exemple: nous avons $N = 4$, $n = 3$, $\mu = 2$, et $\sigma^2 = \frac{7}{2}$. Selon les expressions développées plus haut, nous obtenons effectivement

$$\mathrm{E}[\overline{Y}] = \qquad \text{et} \qquad \mathrm{V}[\overline{Y}] = \qquad .$$

La composante $\frac{N-n}{N-1}$ est le
; sa présence s'explique puisque la population n'est pas infinie.

Comme l'EAS est construit sans remplacer les unités dans la population finie, la présence d'une unité dans l'EAS affecte la probabilité qu'une autre unité s'y retrouve aussi – .

Lorsque N est "large" et que le rapport $\frac{n}{N}$ est "petit", le ≈ 1 et la situation s'apparente à un échantillonnage .

Résumé – EAS – moyenne: (par abus de notation, on écrit y_i, $\overline{y}$, $\mathrm{E}(\overline{y})$, $\mathrm{V}(\overline{y})$, etc.)

- échantillon: $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U} = \{u_1, \ldots, u_N\}$
- moyenne et variance de $\mathcal{U}$: μ, σ^2
- moyenne empirique: $\overline{y} = \frac{1}{n}(y_1 + \cdots + y_n)$
- estimateur sans biais: $\mathrm{E}(\overline{y}) = \mu$
- variance d'échantillonnage: $\mathrm{V}(\overline{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$
- erreur d'estimation: $\mathrm{SD}(\overline{y}) = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$

2.3 – Estimation et intervalles de confiance

L'estimateur $\overline{y}$ est **sans biais d'échantillonnage** sous un EAS. Dans ce scénario, quelle interprétation donner à la variabilité d'échantillonnage $V(\overline{y})$?

Elle donne une idée de la distance typique entre la
et la

L'**erreur quadratique moyenne** de $\overline{y}$ sous un EAS est

$$\mathrm{EQM}(\overline{y}) = \qquad ,$$

c'est-à-dire que l' est entièrement dominée par $\mathrm{V}(\overline{y})$.

Lorsque l'on échantillonne (¬EAS), les observations $y_1, \ldots, y_n$ sont considérées être .

Si de plus elles sont de sorte à ce que $\mathrm{E}(y_i) = \mu$ et $\mathrm{V}(y_i) = \sigma^2$ pour $1 \leq i \leq n$, alors

$$\mathrm{E}(\overline{y}) = \quad , \quad \text{et} \quad \mathrm{V}(\overline{y}) = \quad .$$

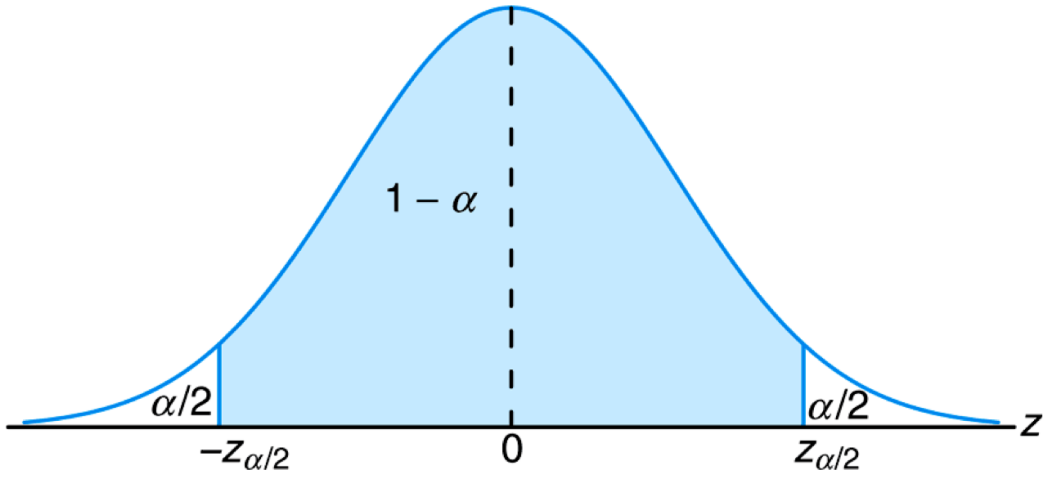
Lorsque $n \to \infty$, $\overline{y} \sim$, selon le TLC, d'où

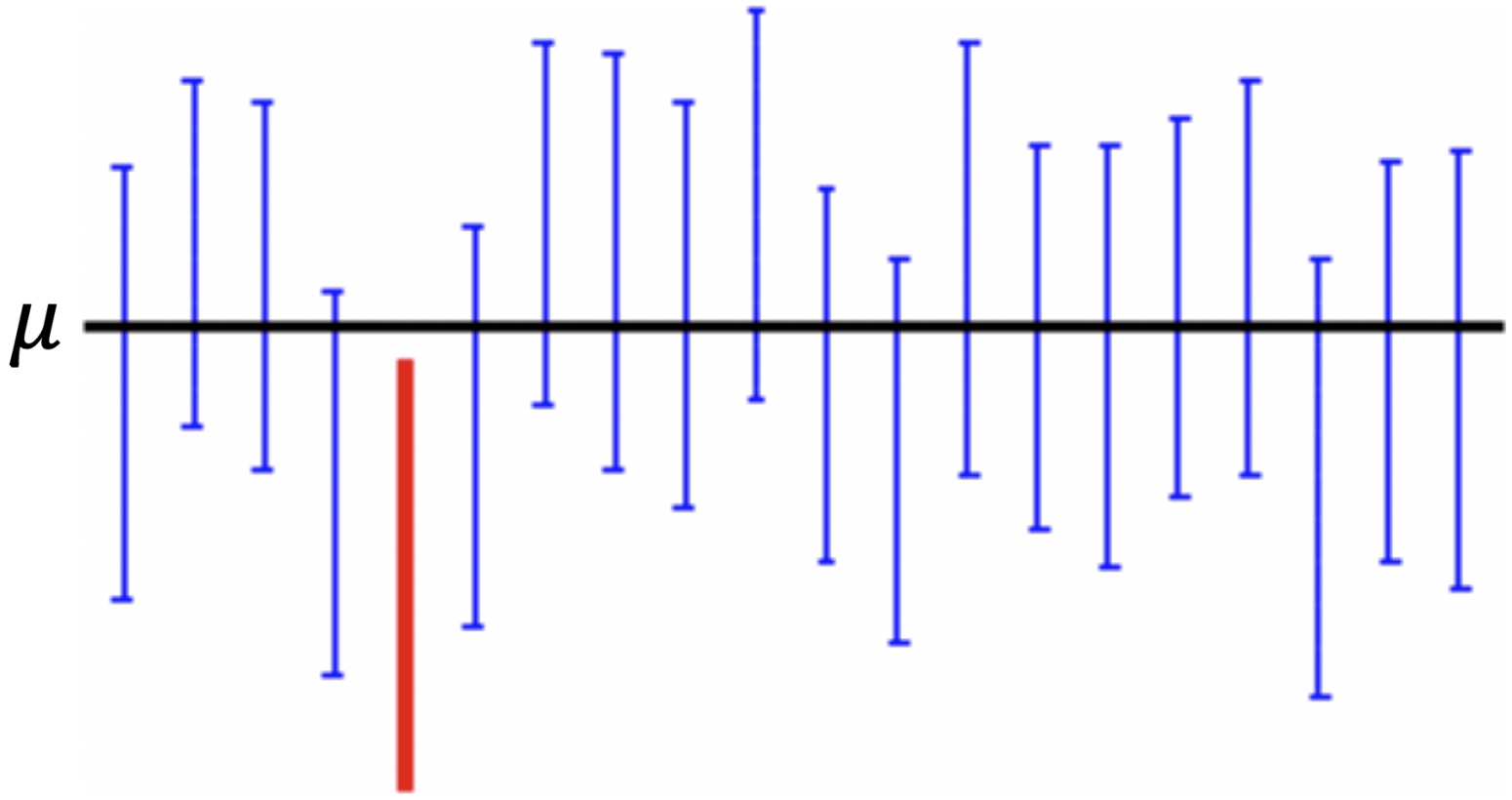
$$Z = \frac{\overline{y} - \mu}{\mathrm{SD}(\overline{y})} = \quad .$$

Soient $\alpha \in (0, 1)$ et $z_\alpha > 0$ le $(1 - \alpha)^{\text{e}}$ quantile d'une variable aléatoire Z suivant la loi normale standard $\mathcal{N}(0, 1)$.

Selon l' de la probabilité, on peut s'attendre à ce que $\frac{\overline{y}-\mu}{\sigma/\sqrt{n}}$ se retrouve dans $(-z_{\alpha/2}, z_{\alpha/2})$ environ $100(1-\alpha)\%$ du temps lorsque l'on répète cette procédure:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{y} - \mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha$$




μ

La quantité

$$B_\alpha = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = z_{\alpha/2}\mathrm{SD}(\overline{y})$$

est une .

On peut alors construire un **intervalle de confiance de** μ

$$\mathsf{IC}(\mu; 100(1-\alpha)\%) : \qquad .$$

Mais nous ne faisons pas affaire à des variables aléatoires indépendantes et identiquement distribuées dans un scénario EAS.

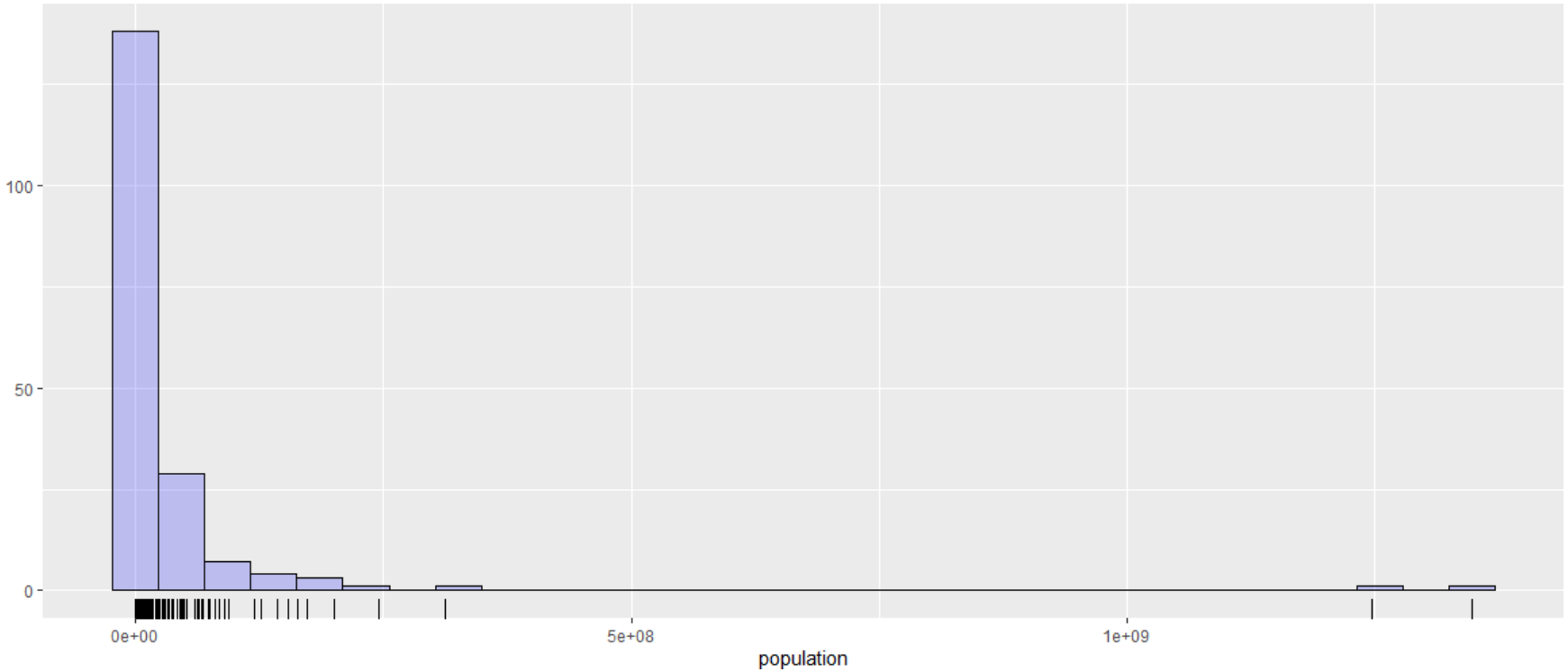
De quelle façon doit on modifier cet argument lorsque les observations proviennent d'une **population finie** et qu'elles sont échantillonnées **sans remise** ?

Nous allons illustrer les principes fondamentaux de la théorie de l'échantillonnage à l'aide de l'ensemble de données `gapminder.csv`, comme on l'a fait au début de ce chapitre.

En plus de la **durée de vie** moyenne, on s'intéressera aussi au **produit national brut** et à la **population** mondiale (en 2011), en particulier à la

- population totale de la planète,
- population moyenne par pays, et
- proportion des pays dont la population est inférieure à 10M.

La population de 185 pays est disponible – elle varie de 56,641 à 1,348,174,478, avec une valeur moyenne de $\mu = 37,080,426$.



La distribution de la population par pays est , avec une queue qui , et deux observations aberrantes ( ).
On retrirera parfois ces observations de l'ensemble de données.

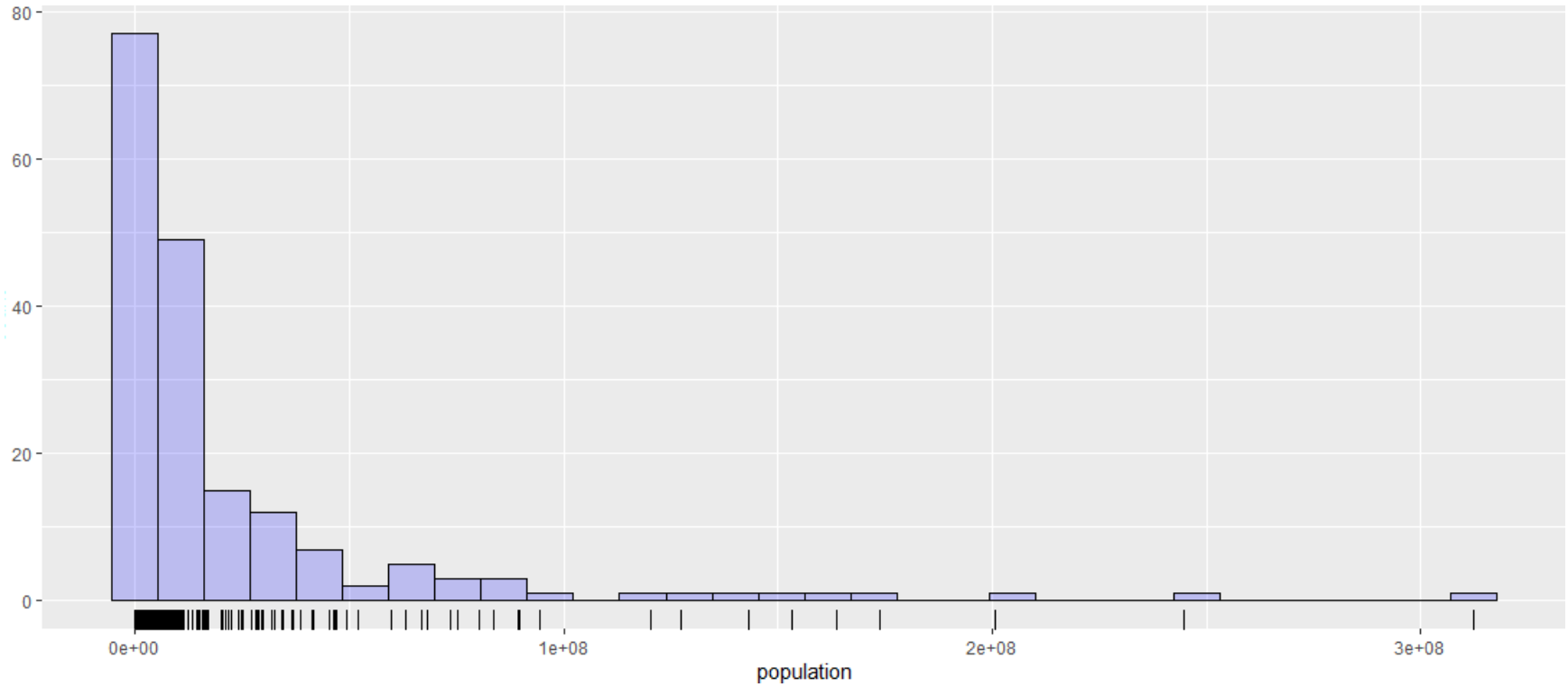
```
# preparation de l'ensemble de donnees
> gapminder.EAS <- gapminder %>%
    filter(year==2011) %>%
    select(population) %>% filter(population < 1000000000)
> summary(gapminder.EAS)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
    56441   2061342   7355231  23301958  22242334 312390368
```

```
> nrow(gapminder.EAS)
```

```
[1] 183
```

```
> ggplot(data=gapminder.EAS, aes(population)) + geom_rug() +
    geom_histogram(col="black", fill="blue", alpha=.2)
```



La distribution associée a la même forme, mais les 183 populations sont inférieures à $312,390,368$, avec une moyenne de $\mu = 23,301,958$.

2.3.1 – Estimation de la moyenne μ

Dans un EAS, nous avons démontré que la moyenne empirique $\overline{y}$ calculée à partir d'un échantillon de taille n est un **estimateur sans biais** de la moyenne μ d'une population de taille N et de variance σ^2.

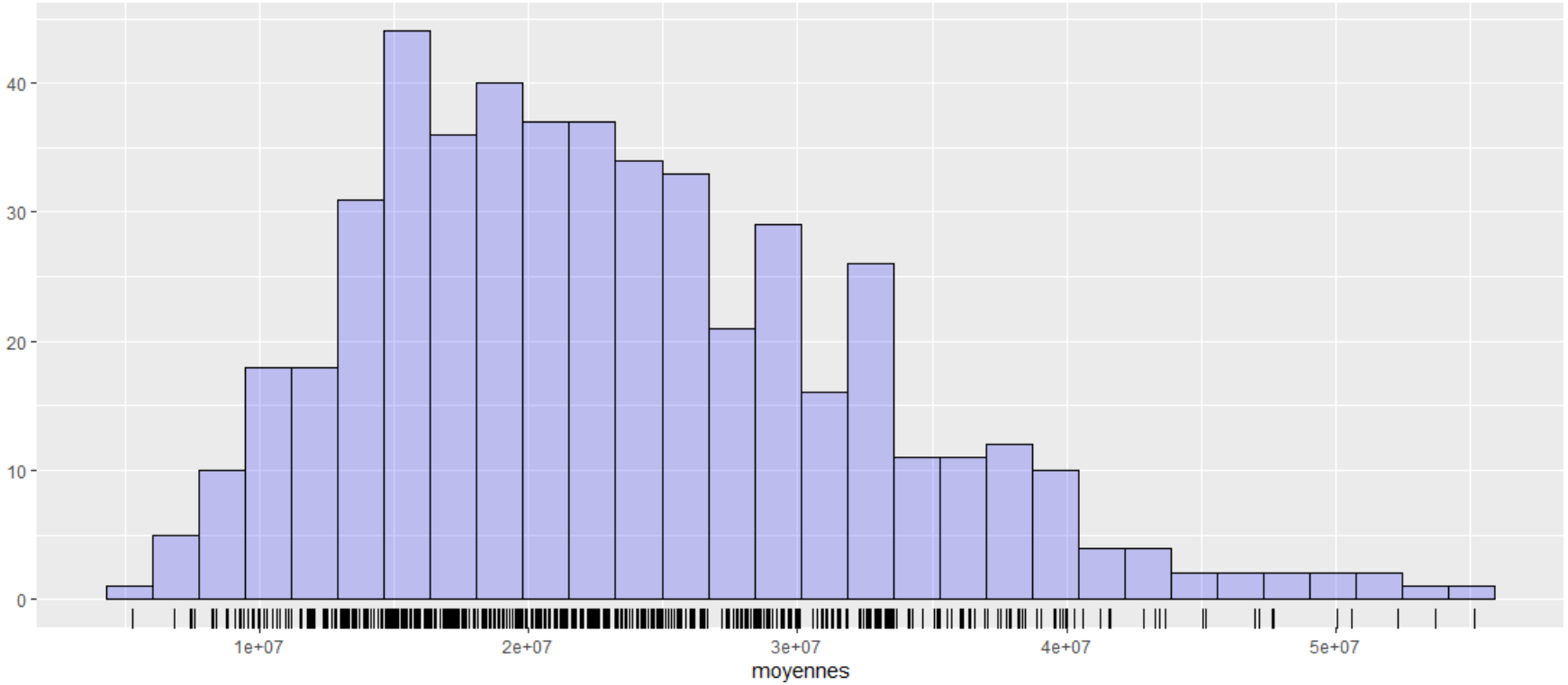
Nous avons également montré que la **variance d'échantillonnage** de l'estimateur $\overline{y}$ est

$$V(\overline{y}) = \frac{\sigma^2}{n}\Big(\frac{N-n}{N-1}\Big).$$

Quelle distribution peut-on s'attendre à ce que $\overline{y}$ suive?

Retournons à l'exemple de la population mondiale.

On produit 500 échantillons de 20 pays choisis selon un EAS à même la liste des 183 pays. Pour chaque échantillon $1 \leq i \leq 500$, on calcule ensuite la **moyenne empirique** $\overline{y}_i$ – leur distribution prend la forme suivante.



```
# 500 echantillons de taille 20
> set.seed(12) # replicabilite
> N=dim(gapminder.EAS)[1]
> n=20
> m=500

# moyennes empiriques
> moyennes <- c()
> for(k in 1:m){
    moyennes[k] <- mean(gapminder.EAS[sample(1:N,n, replace=FALSE),])
  }

>  summary(moyennes)

# histogramme des moyennes empiriques
> ggplot(data=data.frame(moyennes), aes(moyennes)) + geom_rug() +
  geom_histogram(col="black", fill="blue", alpha=.2)
```

Quoique la distribution des moyennes empiriques $\overline{y}_i$ est toujours , la courbe de densité s'apparente tout de même à celle d'une .

Théorème de la limite centrée – EAS
Soient $\mathcal{U} = \{u_1, \ldots, u_N\}$ une population finie de moyenne μ et de variance σ^2, et $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U}$ un **échantillon aléatoire simple**. Si n et $N - n$ sont suffisamment élevés, alors

$$\overline{y} \sim_{\text{approx.}} \quad .$$

Dans un EAS, la **marge d'erreur sur l'estimation** et l'**I.C. à 95%** sont

$$B_\mu = \qquad \text{et} \quad P(|\overline{y} - \mu| \leq B_\mu) \approx \quad .$$

En pratique, la **variance de la population** σ^2 est rarement connue. On l'approxime alors souvent à l'aide de la **variance empirique**

$$s^2 = \qquad\qquad , \quad \{y_i\} \text{ i.i.d.}$$

Mais s^2 est un de σ^2 dans un scénario où l'EAS provient d'une **population finie**. En effet,

$$\mathrm{E}(s^2) =$$

$$=$$

$=$

$=$

$=$ $= \frac{N}{N-1}\sigma^2.$

L'estimateur de σ^2 dans le contexte EAS est ainsi

$\frac{N-1}{N}s^2$ puisque $\mathrm{E}\left[\frac{N-1}{N}s^2\right] =$.

On approxime alors la **variance d'échantillonnage** en substituant σ^2 par $\frac{N-1}{N}s^2$ dans $\mathrm{V}(\overline{y})$:

$$\hat{\mathrm{V}}(\overline{y}) = \qquad\qquad = \frac{s^2}{n}\left(1 - \frac{n}{N}\right).$$

La **marge d'erreur sur l'estimation** est approchée par

$$B_\mu \approx \hat{B}_\mu = 2\sqrt{\hat{\mathrm{V}}(\overline{y})} = \qquad\qquad ,$$

d'où

$$\mathrm{IC}(\mu; 0.95) :$$

forme un **intervalle de confiance de μ à environ 95%**.

Si la variance σ^2 est **connue**, le FCPF est $\frac{N-n}{N-1}$; si la variance σ^2 est **inconnue**, le FCPF est $1 - \frac{n}{N}$. En pratique, lorsque le **taux d'échantillonnage** $\frac{n}{N}$ est inférieur à 5%, on peut laisser tomber le FCPF ( ).

Exemple:
Considérons un échantillon $\mathcal{Y}$ de taille $n = 132$ prélevé à même une population finie $\mathcal{U}$ de taille $N = 37,444$. Supposons que la moyenne et l'écart-type empirique de l'échantillon soient $\overline{y} = 111.3$ et $s = 16.35$, respectivement. Donner un I.C. de la moyenne μ de $\mathcal{U}$ à environ 95%.

Solution:

.

Exemple:
Donner un intervalle de confiance à 95% de la pop. moyenne par pays en 2011 (en excluant l'Inde et la Chine), en utilisant un EAS de taille $n = 20$.

Solution: On doit tout d'abord calculer $\overline{y}$, s^2 (les résultats varient d'un échantillon à l'autre).

```
# IC 95% de mu, n=20
> set.seed(12) # replicabilite
> N = dim(gapminder.EAS)[1]
> n = 20
> (mu = mean(gapminder.EAS[,]))
> sigma.2= mean((gapminder.EAS[,] - mean(gapminder.EAS[,]))^2)
```

```
[1] 23301958
```

La moyenne empirique de l'échantillon est:

```
# echantillon
> EAS = gapminder.EAS[sample(1:N,n, replace=FALSE),]
> (y.barre = mean(EAS))
```

```
[1] 25085501
```

Si on connait la variance, on calcule B et l'I.C. de μ à l'aide de la formule:

```
# variance connue
> B.sigma = 2*sqrt(sigma.2/n*(N-n)/(N-1))
> (I.C.sigma = c(y.barre-B.sigma,y.barre+B.sigma))
```

```
[1]  6759624 43411378
```

Si on ne connait pas la variance, on calcule B et l'I.C. de μ à l'aide de l'autre formule:

```
# variance inconnue
> s.2 = var(EAS)
> borne.erreur.s = 2*sqrt(s.2/n*(1-n/N))
> (I.C.s = c(y.barre-borne.erreur.s,y.barre+borne.erreur.s))
```

```
[1]  6755099 48038164
```

Dans les deux cas, la moyenne actuelle $\mu = 23,301,958$ est contenue dans l'intervalle de confiance.

On remarque de plus que l'I.C. à 95% quand la variance σ^2 est connue est contenu dans l'I.C. à 95% quand la variance ne l'est pas; est-ce toujours le cas?

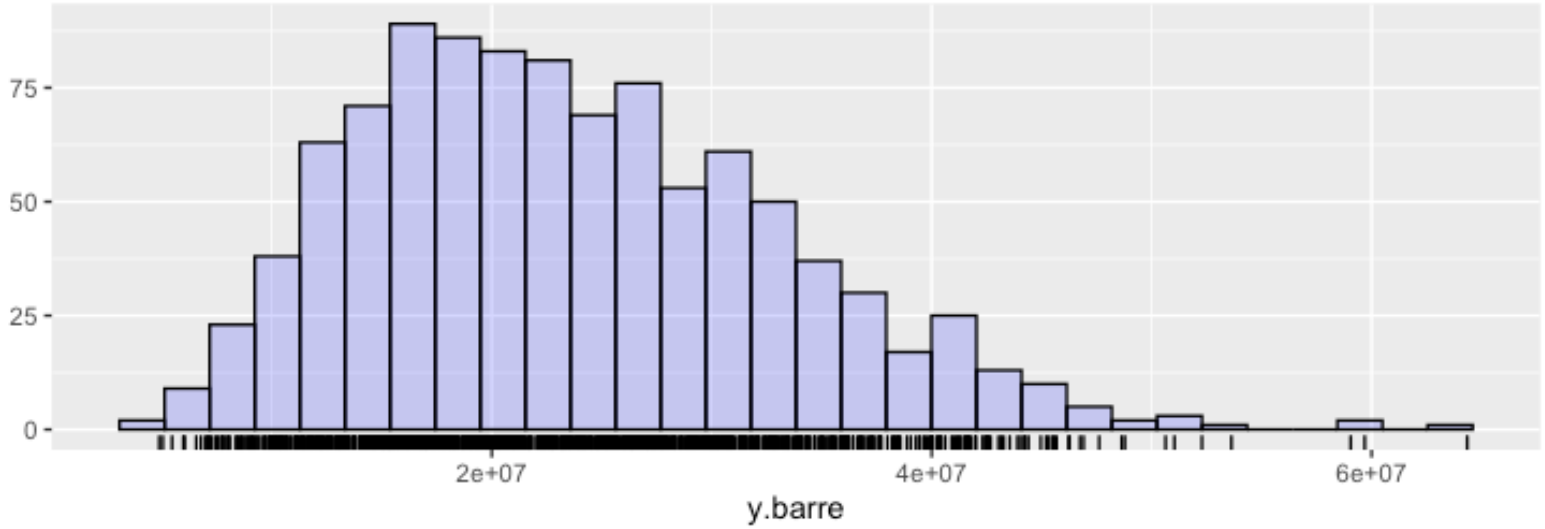
Mais il se pourrait que l'I.C. à 95% construit à partir d'un échantillon ne contienne pas la moyenne μ. En répétant la procédure 1000 fois...

```
# variance inconnue
> m = 1000
> mu.dans.I.C.s = c()
> for(j in 1:m){
    test = gapminder.EAS[sample(1:N,n, replace=FALSE),]
    s.2 = var(test)
    borne.erreur.s = 2*sqrt(s.2/n*(1-n/N))
    mu.dans.I.C.s[j] = mean(test)-borne.erreur.s < mu & mu <
      mean(test)+borne.erreur.s
  }
> mean(mu.dans.I.C.s)
```

```
[1] 0.802
```

Ce n'est pas le $\approx 95\%$ auquel on s'attendait; mais si on passe à des EAS de taille $n = 50, 70, 90$ la proportion se rapproche de 95%.

La longue queue de la distribution de la population pour les $N = 183$ unités y est sans doute pour quelque chose – la distribution des $\overline{y}$ pour $m = 1000$ échantillons de taille $n = 20$ s'éloigne d'une loi normale.



Exemple:
Donner un intervalle de confiance à 95% de l'espérance de vie moyenne par pays en 2011 (en incluant l'Inde et la Chine), en utilisant un EAS de taille $n = 20$.

Solution: On peut ré-utiliser le même code, en modifiant simplement l'ensemble de départ:

```
> gapminder.EAS <- gapminder %>%
   filter(year==2011) %>%
   select(life_expectancy)
```

On obtient alors la moyenne $\mu = 71.18$ et la variance $\sigma^2 = 68.74$. Notre échantillon de taille $n = 20$ a une moyenne empirique de $\overline{y} = 71.66$ (les résultats peuvent changer d'un échantillon à l'autre).

Si la variance est connue on calcule B_μ et l'I.C. de μ à l'aide des formules et on obtient:

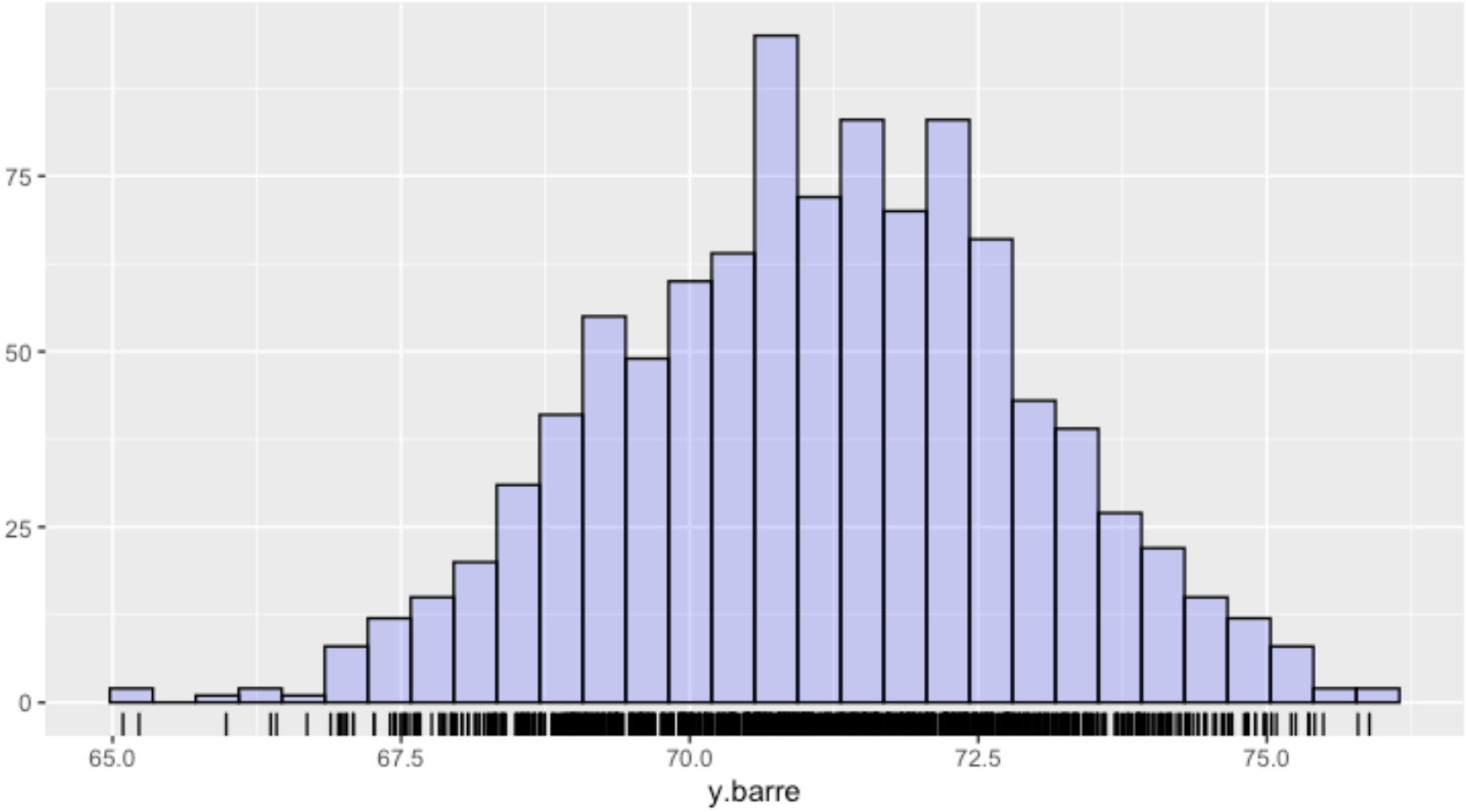
$$\text{IC}(\mu; 0.95): \quad \overline{y} \pm B_\mu \equiv (68.15, 75.18).$$

Si la variance n'est pas connue, on obtient plutôt

$$\text{IC}(\mu; 0.95): \quad \overline{y} \pm \hat{B}_\mu \equiv (69.02, 74.31).$$

Dans les deux cas, la moyenne actuelle μ est contenue dans l'intervalle de confiance.

En répétant la procédure à 1000 reprises, on voit que μ se retrouve dans l'intervalle de confiance environ 92.5% du temps (ce n'est pas tout-à-fait 95%, mais c'est beaucoup mieux qu'à l'exemple précédent).


75
50
25
0
65.0
67.5
70.0
72.5
75.0
y.barre

Résumé – EAS – moyenne: (continuation)

- échantillon: $\mathcal{Y} = \{y_1, \ldots, y_n\} \subseteq \mathcal{U} = \{u_1, \ldots, u_N\}$
- quantités: μ, σ^2, $\overline{y}$, s^2
- borne sur l'erreur d'estimation (σ^2 connue): $B_\mu = 2\sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$
- intervalle de confiance de μ à 95% (σ^2 connue): $\mathsf{IC}(\mu; 0.95) = \overline{y} \pm B_\mu$
- borne sur l'erreur d'estimation (σ^2 inconnue): $\hat{B}_\mu = 2\sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$
- intervalle de confiance de μ à 95% (σ^2 inconnue): $\mathsf{IC}(\mu; 0.95) = \overline{y} \pm \hat{B}_\mu$

2.3.2 – Estimation du total τ

Le gros du travail a déjà été effectué: puisque le **total** τ se ré-écrit

$$\tau = \sum_{j=1}^{N} u_j = N\mu,$$

on peut estimer le total à l'aide d'un EAS en utilisant la formule

$$\hat{\tau} = \qquad .$$

C'est un estimateur non-biaisé du total puisque son **espérance** est

$$\mathrm{E}(\hat{\tau}) = \qquad .$$

Sa **variance d'échantillonnage** s'exprime par

$$\mathrm{V}(\hat{\tau}) = \qquad ,$$

d'où la **marge d'erreur sur l'estimation** est

$$B_\tau = 2\sqrt{\mathrm{V}(\hat{\tau})} = \qquad .$$

Puisqu'en général la variance σ^2 de la population finie $\mathcal{U}$ est inconnue, on l'approxime en substituant σ^2 par la variance empirique s^2 calculée à même l'échantillon, que l'on multiplie par le facteur de correction .

Rappel: s^2 est

.

L'**approximation de la variance d'échantillonnage** est alors

$$\hat{\mathrm{V}}(\hat{\tau}) = \qquad ,$$

d'où l'**approximation de la marge d'erreur sur l'estimation** est

$$B_\tau \approx \hat{B}_\tau = 2\sqrt{\hat{\mathrm{V}}(\hat{\tau})} = \qquad ,$$

et l'**intervalle de confiance approximatif de τ à 95%** est

$$\mathsf{IC}(\tau; 0.95) : \qquad .$$

Exemple:
Considérons un échantillon $\mathcal{Y}$ de taille $n = 132$ prélevé à même une population finie $\mathcal{U}$ de taille $N = 37,444$. Supposons que la moyenne et l'écart-type empirique de l'échantillon soient $\overline{y} = 111.3$ et $s = 16.35$, respectivement. Donner un I.C. du total τ de $\mathcal{U}$ à environ 95%.

Solution:

Exemple:
Donner un intervalle de confiance à 95% de la population de la planète excluant la Chine et l'Inde, en utilisant un EAS de taille $n = 20$.

Solution: on peut se servir de l'échantillon obtenu au préalable, pour lequel nous avons:

$$\overline{y} = 27,396,632 \quad \text{et} \quad \text{IC}(\mu; 0.95) : \quad (6,755,099; 48,038,164).$$

Alors $\hat{B}_\mu \approx$ et

$$\hat{B}_\tau \approx N\hat{B}_\mu = ,$$

d'où $\text{IC}(\tau; 0.95) : N\overline{y} \pm B_\tau =$, ou

$\text{IC}(\tau; 0.95) :$ ... c'est vrai, mais...

2.3.3 – Estimation d'une proportion p

Dans une population où $u_j \in \{0, 1\}$ représente une **réponse binaire** pour tout $1 \leq j \leq N$ (par exemple, $u_j = 1$ quand l'unité correspondante possède une certaine caractéristique), la **moyenne** prend une interprétation particulière:

$$p = \mu = \frac{1}{N} \sum_{j=1}^{N} u_j$$

est la .

On peut estimer cette proportion à l'aide d'un EAS en utilisant la formule

$$\hat{p} = \overline{y} = .$$

C'est un estimateur non-biaisé de la proportion puisque son **espérance** est

$$\mathrm{E}(\hat{p}) = \qquad .$$

Sa **variance d'échantillonnage** s'exprime par

$$\mathrm{V}(\hat{p}) = \qquad .$$

Mais $U^2 = U$ lorsque la réponse U est binaire, et ainsi

$$\sigma^2 =$$

d'où

$$\mathrm{V}(\hat{p}) = \qquad .$$

La **marge d'erreur sur l'estimation** est

$$B_p = 2\sqrt{\mathrm{V}(\hat{p})} = \qquad .$$

Quand la variance σ^2 de la population finie $\mathcal{U}$ est inconnue (c-à-d que p est inconnue), l'**approximation de la variance d'échantillonnage** est alors

$$\hat{\mathrm{V}}(\hat{p}) = \qquad .$$

Lorsque la réponse y_i est binaire, alors $y_i^2 = y_i$ pour tout $1 \leq i \leq n$ et

$$s^2 = \frac{1}{n-1}\Big(\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\Big) = \qquad ,$$

d'où

$$\hat{\mathrm{V}}(\hat{p}) = \qquad .$$

L'**approximation de la marge d'erreur sur l'estimation** est

$$B_p \approx \hat{B}_p = 2\sqrt{\hat{\mathrm{V}}(\hat{p})} = \qquad ,$$

et l'**intervalle de confiance approximatif de p à 95%** est

$$\mathsf{IC}(p; 0.95) : \qquad .$$

Exemple:
Considérons un échantillon $\mathcal{Y}$ de taille $n = 132$ prélevé à même une population finie $\mathcal{U}$ de taille $N = 37,444$. Supposons que 25 des observations de $\mathcal{Y}$ possèdent une caractéristique particulière. Donner un I.C. de la proportion p des observations de $\mathcal{U}$ qui possèdent la caractéristique, à environ 95%.

Solution:

Exemple:
Donner un intervalle de confiance à 95% de la proportion des pays dont la population se retrouve sous le seuil de 10M à l'aide d'un EAS, $n = 20$.

Solution: On doit tout d'abord calculer $\hat{p}$ (les résultats varient d'un échantillon à l'autre).

```
# ensemble de donnees
> gapminder.EAS <- gapminder %>%
    filter(year==2011) %>% select(population)

# IC 95% de p, n=20
> set.seed(1234) # replicabilite (mac)
> N=dim(gapminder.EAS)[1]
> n=20
```

La proportion réelle parmis les $N = 185$ pays est:

```
# creer une variable binaire
> seuil.10 <- gapminder.EAS < 10000000
>  (p = mean(seuil.10))
```

```
[1] 0.5675676
```

La proportion dans l'échantillon de taille $n = 20$ est:

```
# echantillon de taille n
> EAS = seuil.10[sample(1:N,n, replace=FALSE)]
> (p.hat = mean(EAS))
```

```
[1] 0.65
```

Si on suppose que la variance est inconnue, la borne $\hat{B}_p$ sur l'erreur d'estimation et l'intervalle de confiance $\text{IC}(p; 0.95)$ sont données par les formules suivantes:

```
# variance inconnue
> B.p = 2*sqrt(p.hat*(1-p.hat)/(n-1)*(1-n/N))
> (I.C.sigma = c(p.hat-B.p,p.hat+B.p))
```

```
[1] 0.4433193 0.8566807
```

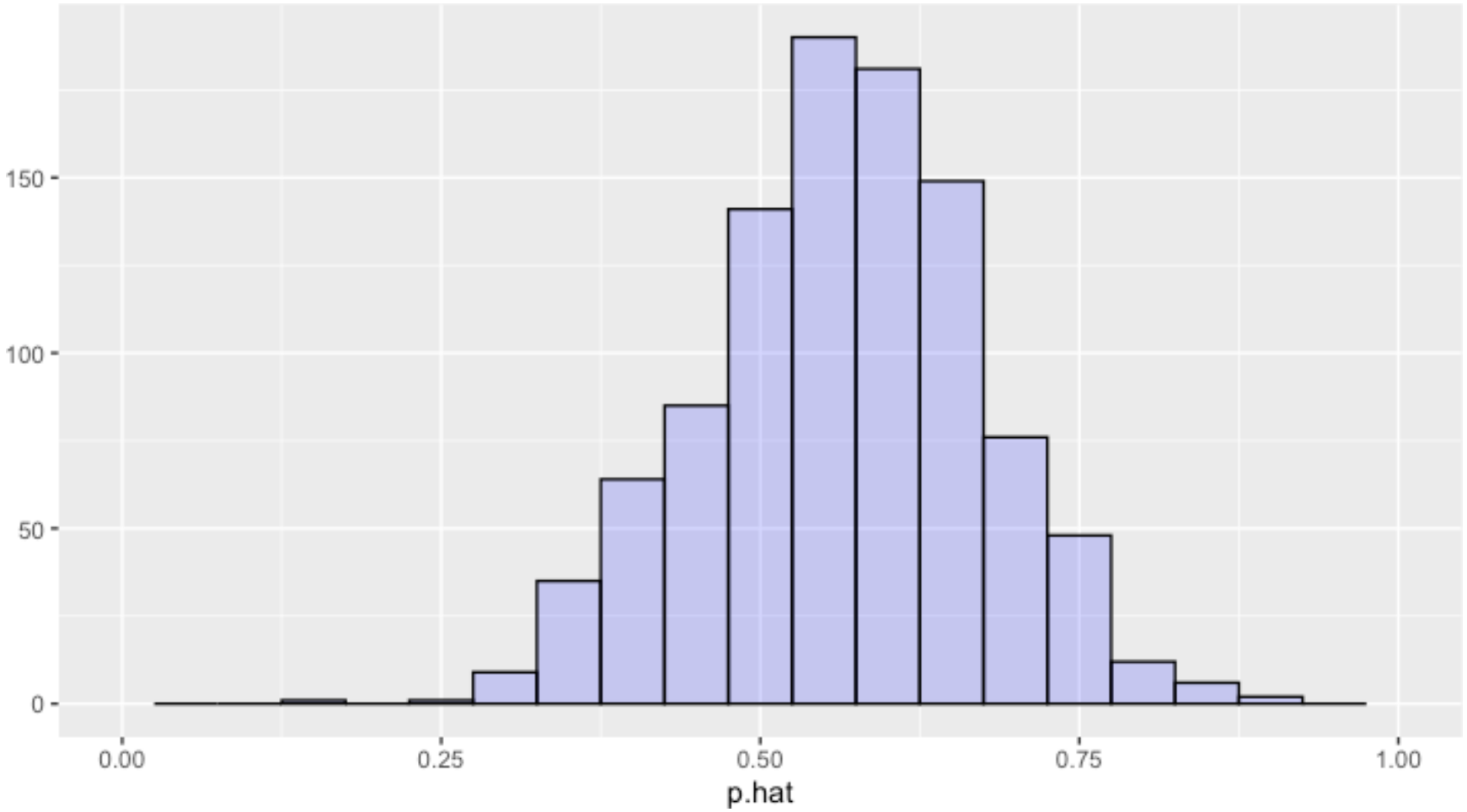
La proportion rélle $p \approx 0.568$ se retrouve effectivement dans l'intervalle de confiance.

On répète la procédure à 1000 reprises; la proportion réelle se retrouve dans l'intervalle de confiance obtenu...

```
> m=1000
> p.dans.I.C.s = c()
> p.hat = c()
> for(j in 1:m){
    p.hat[j] = mean(seuil.10[sample(1:N,n, replace=FALSE)])
    B.p = 2*sqrt(p.hat[j]*(1-p.hat[j])/(n-1)*(1-n/N))
    p.dans.I.C.s[j] = p.hat[j]-B.p < p & p < p.hat[j]+B.p
  }
> mean(p.dans.I.C.s)

> ggplot(data=data.frame(p.hat), aes(p.hat)) +
    geom_histogram(bins=21, col="black", fill="blue", alpha=.2) +
    xlim(0,1)
```

```
[1] 0.934
```

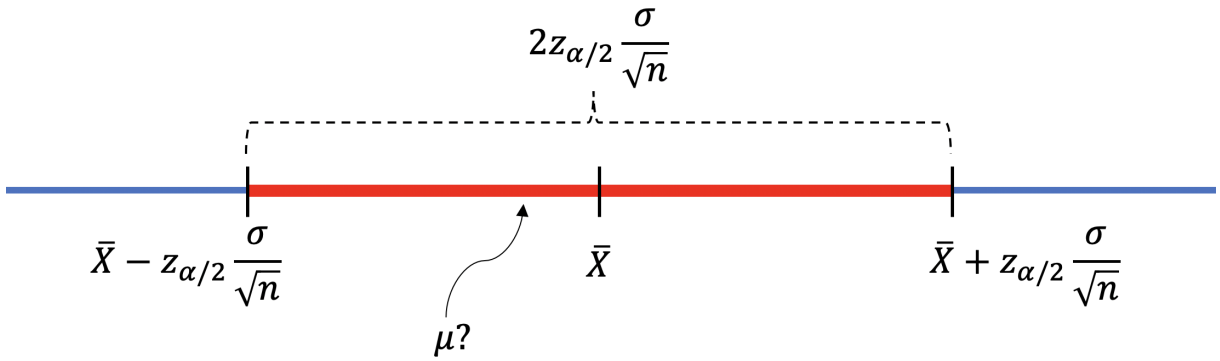

150
100
50
0
0.00
0.25
0.50
0.75
1.00
p.hat

2.4 – Taille de l'échantillon

Lorsque l'on prélève un échantillon $\mathcal{Y}$ de taille n (¬EAS) de $\mathcal{U}$, l'erreur que l'on commet en estimant la moyenne μ à l'aide de la moyenne empirique $\overline{X}$ est en général bornée par

$$B_\alpha = z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

à un niveau de confiance de $100(1-\alpha)\%$.



À un niveau de confiance de 95%, on utilise souvent $B_{0.05} \approx B =$.

Afin de contrôler l'erreur, il faut **contrôler la taille de l'échantillon**:

$$B > \frac{2\sigma}{\sqrt{n}} \implies n > \qquad .$$

Exemple: On prélève un échantillon avec remise d'une population dont la variance est $\sigma^2 = 100$. Quelle est la taille d'échantillon requise afin de s'assurer que la marge d'erreur de la moyenne, à un niveau de confiance d'environ 95%, soit inférieure à 1.2?

Solution: Il suffit d'utiliser

$$n > \qquad .$$

Mais ce n'est pas de cette manière que cela se déroule dans un sondage.

En premier lieu, il y a un problème **pratique** lié à l'échantillonnage.

Puisque le coût associé à chaque réponse peut s'avérer **dispendieux** (en termes de /), on cherche souvent à minimiser la taille de l'**échantillon** $\mathcal{Y}$ étant donnée une **marge d'erreur visée**.

En second lieu, la marge d'erreur (visée) dans un EAS s'exprime selon

$$B_\xi = 2\sqrt{\mathrm{V}(\hat{\xi})}, \quad \xi \in \{\mu, \tau, p\}.$$

Mais ces variances dépendent toutes de la taille n de l'échantillon $\mathcal{Y}$.

On cherche ainsi à exprimer n en termes des paramètres N, σ^2, et B_ξ.

2.4.1 – Moyenne μ

Lorsque l'on cherche à estimer μ, nous obtenons

$$B_\mu = 2\sqrt{\frac{\sigma^2}{n}\Big(\frac{N-n}{N-1}\Big)} \iff$$

$$\iff$$

$$\iff$$

$$\iff n_\mu = \frac{N\sigma^2}{(N-1)D_\mu + \sigma^2}.$$

Évidemment, on ne peut utiliser cette formule que **si l'on connaît la variance** σ^2 de la population $\mathcal{U}$ à l'étude.

Il vous vient peut-être l'idée d'utiliser la **variance empirique** s^2 de l'échantillon $\mathcal{Y}$ comme nous l'avons fait pour estimer la variance d'échantillonnage, mais ...

.

Stratagèmes (pour obtenir σ^2):

- prélever un **échantillon préliminaire** ,
- utiliser la variance empirique d'un sondage préalable, ou
- pour une proportion, utiliser un estimé conservateur ($p = 0.5$).

Exemple:
Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$. Supposons que l'on cherche à estimer la moyenne μ de la variable réponse de $\mathcal{U}$. Dans un EAS préliminaire de taille $n = 132$, on calcule un écart-type (empirique) de $s = 16.35$.

En utilisant $\sigma = s$, déterminer la taille requise n_μ d'un échantillon $\mathcal{Y}$ permettant d'estimer la moyenne de la réponse avec une marge d'erreur (visée) d'au plus $B_\mu = 1.7$.

Solution: Il suffit d'utiliser les formules:

$$D_\mu = \qquad\qquad \Longrightarrow \; n_\mu = \qquad\qquad .$$

2.4.2 – Total τ

Lorsque l'on cherche à estimer τ, nous obtenons

$$B_\tau = 2\sqrt{N^2 \cdot \frac{\sigma^2}{n}\Big(\frac{N-n}{N-1}\Big)} \iff$$

$$\iff$$

$$\iff$$

$$\iff n_\tau = \frac{N\sigma^2}{(N-1)D_\tau + \sigma^2}.$$

Exemple:
Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$. Supposons que l'on cherche à estimer le total τ de la variable réponse de $\mathcal{U}$. Dans un EAS préliminaire de taille $n = 132$, on calcule un écart-type (empirique) de $s = 16.35$.

En utilisant $\sigma = s$, déterminer la taille requise n_τ d'un échantillon $\mathcal{Y}$ permettant d'estimer le total de la réponse avec une marge d'erreur (visée) d'au plus $B_\tau = 10000$.

Solution: Il suffit d'utiliser les formules:

$$D_\tau = \qquad\qquad \Longrightarrow \; n_\tau = \qquad\qquad .$$

2.4.3 – Proportion p

Lorsque l'on cherche à estimer p, nous obtenons

$$B_p = 2\sqrt{\frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)} \iff$$

$$\iff$$

$$\iff$$

$$\iff n_p = \frac{Np(1-p)}{(N-1)D_p + p(1-p)}.$$

Exemple:
Considérons une population finie $\mathcal{U}$ de taille $N = 37,444$. Supposons que l'on cherche à estimer la proportion p des unités qui possèdent une caractéristique particulière. Dans un EAS préliminaire de taille $n = 132$, on identifie 25 observations possédant la caractéristique.

En utilisant l'approximation $\sigma^2 = \frac{25}{132} \cdot \frac{107}{132}$ provenant de l'échantillon préliminaire, déterminer la taille requise n_p d'un échantillon $\mathcal{Y}$ permettant d'estimer la proportion de la réponse possédant la caractéristique avec une marge d'erreur (visée) d'au plus $B_p = 0.03$.

Solution: Il suffit d'utiliser les formules:

$$D_p = \qquad\qquad \Longrightarrow \; n_p = \qquad\qquad .$$

Exemple:
Considérons une situation semblable à celle de l'exemple précédent.

En utilisant l'approximation $\sigma^2 = (0.5)^2$, déterminer la taille requise n_p d'un échantillon $\mathcal{Y}$ permettant d'estimer la proportion de la réponse possédant la caractéristique avec une marge d'erreur (visée) d'au plus $B_p = 0.03$.

Solution: Il suffit d'utiliser les formules:

$$D_p = \qquad\qquad \Longrightarrow \; n_p = \qquad\qquad .$$