

# **MAT 3777**

## **Échantillonnage et sondages**

### **Chapitre 4**

### **Estimation par le quotient, par la régression, et par la différence**

P. Boily (uOttawa)

Session d'hiver – 2022

P. Boily (uOttawa)

## Aperçu

### 4.1 – Motivation (p.3)

### 4.2 – Estimation par le quotient (p.5)

- Estimateur du quotient (p.6)
- Biais de l'estimateur du quotient (p.9)
- Variabilité de l'estimateur du quotient (p.16)
- Intervalle de confiance de l'estimateur du quotient (p.23)
- Estimation de la moyenne et du total (p.30)
- Taille de l'échantillon (p.36)

### 4.3 – Estimation par la régression (p.41)

- Estimateur de régression (p.43)
- Biais de l'estimateur de régression (p.48)

- Variabilité de l'estimateur par la régression (p.51)
- Intervalle de confiance de l'estimateur par la régression (p.52)
- Taille de l'échantillon (p.62)

#### 4.4 – Estimation par la différence (p.67)

#### 4.5 – Comparaisons (p.81)

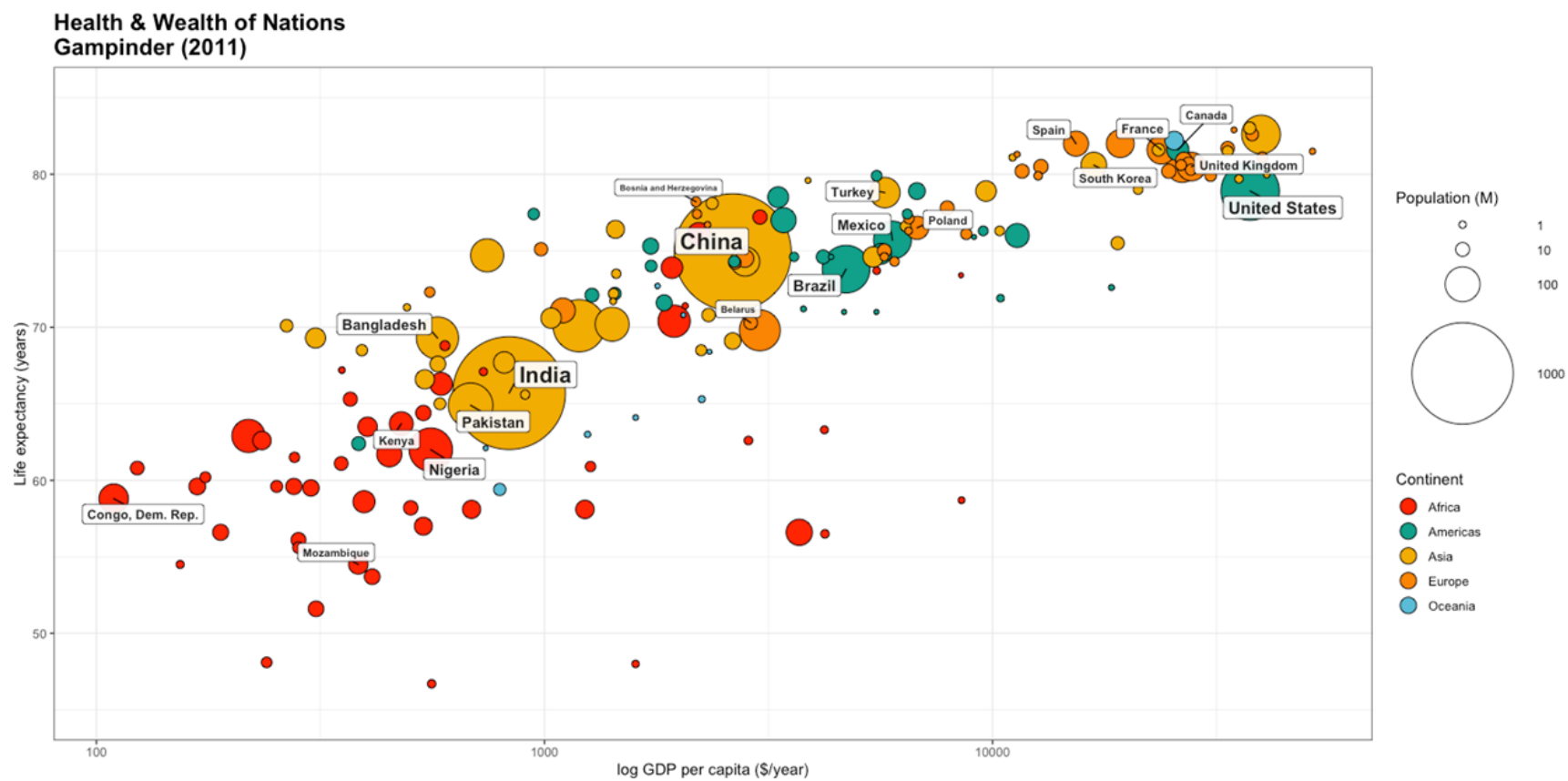
- ... entre EAS et la méthode du quotient (p.82)
- ... entre EAS et la méthode de la régression (p.87)
- ... entre EAS et la méthode de la différence (p.88)
- ... entre les trois méthodes (p.89)

## 4.1 – Motivation

Jusqu'à présent, nous avons discuté d'estimateurs EAS et STR **univarés**. Peut-on utiliser plus d'une réponse par unité afin d'obtenir de meilleures approximations (c'est-à-dire des estimateurs possédant des variances plus faibles, en termes relatifs).

Dans l'ensemble de données `gapminder.csv`, il y a  $N = 168$  observations pour lesquelles l'**espérance de vie**  $Y$  et le (logarithme du) **produit national brut par habitant**  $X$  des pays de la planète en 2011 sont disponibles.

Supposons qu'il soit connu que  $E[X] = \mu_X = 7.84$ . Si on préleve un échantillon  $\{(x_1, y_1), \dots, (x_{10}, y_{10})\} \subseteq \mathcal{U}$  pour lequel la moyenne des  $y_i/x_i$  est 8.67, peut-on s'attendre à ce que  $\mu_Y \approx 8.67\mu_X = 68.00$ ?



## 4.2 – Estimation par le quotient

Soit  $\mathcal{U} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$  une population finie de taille  $N$  pour laquelle chaque unité  $u_j$  admet 2 réponses:  $X_j$  et  $Y_j$ .

Le **quotient des moyennes**  $R$  est le rapport des moyennes (ou des totaux):

$$R = \frac{\sum_{j=1}^N Y_j}{\sum_{j=1}^N X_j} = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}, \quad \text{tant que } \mu_X, \tau_X \neq 0.$$

On s'intéresse à de tels quotients lorsque l'on cherche à déterminer le salaire moyen  $Y$  en fonction de l'âge  $X$  au Canada, par exemple.

### 4.2.1 – Estimateur du quotient

Soit  $\mathcal{Y} = \{(x_{i_1}, y_{i_1}), \dots, (x_{i_n}, y_{i_n})\} \subseteq \mathcal{U}$  un **échantillon aléatoire simple bivarié** de taille  $n$ . On allège souvent la notation en écrivant

$$\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}.$$

Le **quotient des moyennes  $r$  de l'échantillon** est un estimateur de  $R$ :

$$r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{\tau}_Y}{\hat{\tau}_X}, \quad \text{tant que } \bar{x}, \hat{\tau}_X \neq 0.$$

Mais c'est un estimateur **biaisé!**

**Exemple:** Soit une population finie à  $N = 4$  éléments et 2 réponses:

$$u_1 = (1, 2), \quad u_2 = (1, 0), \quad u_3 = (2, 1), \quad u_4 = (4, 5).$$

La quotient des moyennes  $R$  est

$$R = \frac{2 + 0 + 1 + 5}{1 + 1 + 2 + 4} = \frac{8}{8} = 1.$$

Supposons que l'on souhaite prélever de cette population un EAS sans remise de taille  $n = 3$  afin d'estimer le quotient  $R$ .

Il y a  $\binom{4}{3} = 4$  tels échantillons.



Échantillon	Valeurs $y$	$\bar{y}$	Valeurs $x$	$\bar{x}$	$r$	$P(r)$
$u_1, u_2, u_3$	2, 0, 1	1	1, 1, 2	4/3	3/4	1/4
$u_1, u_2, u_4$	2, 0, 5	7/3	1, 1, 4	2	7/6	1/4
$u_1, u_3, u_4$	2, 1, 5	8/3	1, 2, 4	7/3	8/7	1/4
$u_2, u_3, u_4$	0, 1, 5	2	1, 2, 3	2	1	1/4

Alors

$$E[r] = \sum_r rP(r) = \frac{1}{4} (3/4 + 7/6 + 8/7 + 1) = \frac{341}{336} \approx 1.014881 \neq R = 1.$$

Le biais d'échantillonnage de  $r$  en tant qu'estimateur de  $R$  est

$$E[r - R] = E \left[ \frac{\bar{y}}{\bar{x}} - R \right] = E \left[ \frac{1}{\bar{x}} (\bar{y} - R\bar{x}) \right] = ??$$

## 4.2.2 – Biais de l'estimateur du quotient

Dans cette expression pour le biais d'échantillonnage, il n'y a que  $\bar{x}$  et  $\bar{y}$  qui changent lorsque l'échantillon change:  $R$  demeure constant.

Mais nous n'avons pas d'expression simple nous permettant de calculer **l'espérance d'un quotient de variables aléatoires**.

Soit  $f : [a, b] \rightarrow \mathbb{R}$  une fonction  $C^2$  (i.e.  $f, f', f''$  sont continues). Selon le théorème de Taylor, pour tout  $c \in (a, b)$ , il existe un  $\xi$  entre  $c$  et  $z$  tel que

$$f(z) = f(c) + f'(c)(z - c) + \frac{f''(\xi)}{2}(z - c)^2.$$

Puisque  $f''$  est continue sur  $[a, b]$ ,  $f''$  est bornée sur  $[a, b]$ :  $\exists M > 0$  tel que  $|f''(z)| \leq M$  pour tout  $z \in [a, b]$ .

Ainsi, si  $z$  est **suffisamment près** de  $c$ ,

$$|f(c) + f'(c)(z - c)| \gg \frac{M}{2}(z - c)^2 \geq \left| \frac{f''(\xi)}{2}(z - c)^2 \right|,$$

d'où

$$f(z) \approx f(c) + f'(c)(z - c).$$

Posons  $f(z) = \frac{1}{z}$ ,  $z = \bar{x}$  et  $c = \mu_X$ . Alors  $f'(z) = -\frac{1}{z^2}$ .

Puisque  $f$  est  $C^2$  sur tout intervalle  $[a, b]$  tel que  $a > 0$ , si  $\bar{x}$  est suffisamment près de  $\mu_X$ , alors

$$\frac{1}{\bar{x}} \approx \frac{1}{\mu_X} - \frac{1}{\mu_X^2}(\bar{x} - \mu_X)$$

(l'approximation constante est  $\frac{1}{\bar{x}} \approx \frac{1}{\mu_X}$ ).

Puisque  $E(\bar{x}) = \mu_X$ ,  $E(\bar{y}) = \mu_Y$  (EAS), et  $\mu_Y = R\mu_X$ , nous obtenons

$$\begin{aligned}
 E[r - R] &= E\left[\frac{\bar{y} - R\bar{x}}{\bar{x}}\right] \approx E\left[\left(\frac{1}{\mu_X} - \frac{1}{\mu_X^2}(\bar{x} - \mu_X)\right)(\bar{y} - R\bar{x})\right] \\
 &= E\left[\frac{1}{\mu_X}(\bar{y} - R\bar{x})\right] - E\left[\frac{1}{\mu_X^2}(\bar{x} - \mu_X)(\bar{y} - R\bar{x})\right] \\
 &= \frac{1}{\mu_X}\left(E(\bar{y}) - R \cdot E(\bar{x})\right) - \frac{1}{\mu_X^2}\left(E[\bar{x}\bar{y}] - \mu_X\bar{y} - R\bar{x}^2 - R\mu_X\bar{x}\right) \\
 &= \frac{1}{\mu_X}\underbrace{(\mu_Y - R\mu_X)}_{=0} - \frac{1}{\mu_X^2}\left(E(\bar{x}\bar{y}) - \mu_X E(\bar{y}) - R(E(\bar{x}^2) - \mu_X E(\bar{x}))\right) \\
 &= -\frac{1}{\mu_X^2}\left(E(\bar{x}\bar{y}) - \mu_X\mu_Y - R(E(\bar{x}^2) - \mu_X^2)\right)
 \end{aligned}$$

Nous simplifions encore le biais d'échantillonnage  $E[r - R]$  à l'aide de  $E(\overline{xy}) = \mu_X \mu_Y + \text{Cov}(\overline{x}, \overline{y})$ , et  $E(\overline{x}^2) = \mu_X^2 + V(\overline{x})$ . Ainsi,

$$E[r - R] \approx -\frac{1}{\mu_X^2} [\text{Cov}(\overline{x}, \overline{y}) - R \cdot V(\overline{x})]$$

Dans un EAS de taille  $n$ , prélevé à même une population finie de taille  $N$  et de variance  $\sigma^2$ , nous avons déjà vu que

$$V(\overline{x}) = \frac{\sigma_X^2}{n} \left( \frac{N-n}{N-1} \right) \quad \text{et} \quad V(\overline{y}) = \frac{\sigma_Y^2}{n} \left( \frac{N-n}{N-1} \right).$$

Considérons la v.a.  $Z = X + Y$ . L'estimateur EAS de  $\mu_Z = \mu_X + \mu_Y$  est

$$\overline{z} = \overline{x} + \overline{y}$$

et sa variance d'échantillonnage est

$$V(\bar{z}) = \frac{\sigma_Z^2}{n} \left( \frac{N-n}{N-1} \right), \quad \text{où}$$

$$\begin{aligned} \sigma_Z^2 &= \frac{1}{N} \sum_{j=1}^N (z_j - \mu_Z)^2 = \frac{1}{N} \sum_{j=1}^N \{ (x_j + y_j) - (\mu_X + \mu_Y) \}^2 \\ &= \frac{1}{N} \sum_{j=1}^N (x_j - \mu_X)^2 + \frac{2}{N} \sum_{j=1}^N (x_j - \mu_X)(y_j - \mu_Y) + \frac{1}{N} \sum_{j=1}^N (y_j - \mu_Y)^2 \\ &= \sigma_X^2 + 2\sigma_{XY} + \sigma_Y^2 = \sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2, \end{aligned}$$

si  $\rho = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y}$  est le **coefficient de corrélation** entre  $X$  et  $Y$ .

D'une part,

$$V(\bar{z}) = \frac{\sigma_X^2 + 2\rho\sigma_X\sigma_Y + \sigma_Y^2}{n} \left( \frac{N-n}{N-1} \right),$$

d'une autre part

$$\begin{aligned} V(\bar{z}) &= V(\bar{x} + \bar{y}) = V(\bar{x}) + 2\text{Cov}(\bar{x}, \bar{y}) + V(\bar{y}) \\ &= \frac{\sigma_X^2}{n} \left( \frac{N-n}{N-1} \right) + 2\text{Cov}(\bar{x}, \bar{y}) + \frac{\sigma_Y^2}{n} \left( \frac{N-n}{N-1} \right); \end{aligned}$$

nous en concluons que

$$\text{Cov}(\bar{x}, \bar{y}) = \frac{\rho\sigma_X\sigma_Y}{n} \left( \frac{N-n}{N-1} \right).$$

Par conséquent,

$$\begin{aligned} E[r - R] &\approx -\frac{1}{\mu_X^2} [\text{Cov}(\bar{x}, \bar{y}) - R \cdot V(\bar{x})] \\ &= -\frac{1}{\mu_X^2} \left[ \frac{\rho \sigma_X \sigma_Y}{n} \left( \frac{N-n}{N-1} \right) - R \frac{\sigma_X^2}{n} \left( \frac{N-n}{N-1} \right) \right] \\ &= \frac{1}{\mu_X^2} \cdot \frac{R \sigma_X^2 - \rho \sigma_X \sigma_Y}{n} \left( \frac{N-n}{N-1} \right) \end{aligned}$$

Mais l'erreur systématique n'est pas la seule manière de qualifier la magnitude de l'erreur commise en utilisant  $r$  afin d'estimer  $R$ : l'**erreur quadratique moyenne** (EQM) de  $r$  est

$$\text{EQM}(r) = E((r - R)^2) = V(r) + (E(r) - R)^2.$$



### 4.2.3 – Variabilité de l'estimateur du quotient

On obtient une approximation de  $V(r)$  en utilisant l'approximation de Taylor d'ordre 0:

$$\frac{1}{\bar{x}} \approx \frac{1}{\mu_X}.$$

Ainsi,

$$V(r) = V(r - R) = V\left[\frac{\bar{y}}{\bar{x}} - R\right] = V\left[\frac{\bar{y} - R\bar{x}}{\bar{x}}\right] \approx V\left[\frac{\bar{y} - R\bar{x}}{\mu_X}\right].$$

Considérons la variable aléatoire  $W = Y - RX$ . Puisque  $\mu_Y = R\mu_X$ ,

$$\mu_W = \mu_Y - R\mu_X = 0.$$

La moyenne de  $W$  dans l'échantillon EAS  $\mathcal{Y}$  est alors

$$\bar{w} = \bar{y} - R\bar{x} \implies V(r) \approx V\left[\frac{\bar{w}}{\mu_X}\right] = \frac{1}{\mu_X^2} V(\bar{w}) = \frac{1}{\mu_X^2} \cdot \frac{\sigma_W^2}{n} \left(\frac{N-n}{N-1}\right),$$

où

$$\sigma_W^2 = \frac{1}{N} \sum_{j=1}^N (W_j - \mu_W)^2 = \frac{1}{N} \sum_{j=1}^N W_j^2 = \frac{1}{N} \sum_{j=1}^N (Y_j - RX_j)^2.$$

Ainsi

$$V(r) \approx \frac{1}{\mu_X^2} \cdot \frac{1}{n} \cdot \frac{1}{N} \sum_{j=1}^N (Y_j - RX_j)^2 \left(\frac{N-n}{N-1}\right).$$

Le rapport entre l'erreur systématique  $E[r - R]$  et l'écart type de  $r$  est alors

$$\frac{E[r - R]}{ET(r)} \approx \frac{1}{\sqrt{n}} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{\sigma_W} \sqrt{\frac{N-1}{N-n}};$$

ainsi lorsque  $n, N \rightarrow \infty$  (en supposant que  $N \gg n$ ),

$$\frac{E[r - R]}{ET(r)} \rightarrow 0.$$

Autrement dit, quoiqu'il soit impossible de se débarrasser du biais, l'erreur d'estimation

$$EQM(r) = V(r) + (E(r) - R)^2$$

est dominée par la variance  $V(r)$  si  $n$  est **suffisamment élevée**.

**Exemple:**

La liste des pays pour lesquels et l'espérance de vie et le (logarithme du) produit national brut par habitant sont disponibles en 2011 contient  $N = 168$  observations.

---

```
> library(tidyverse)
> gapminder.QRD <- gapminder %>% filter(year==2011) %>%
  select(life_expectancy,gdp,population)

# on ne garde que les observations qui ont les deux
> gapminder.QRD <- gapminder.QRD[complete.cases(gapminder.QRD),]
> gapminder.QRD <- gapminder.QRD %>% mutate(lgdppc=log(gdp/population))
> (N=nrow(gapminder.QRD))
```

---

[1] 168

On produit 500 échantillons de 20 pays choisis selon un EAS, et on calcule l'estimateur  $r$  du quotient  $R$  pour chacun de ces échantillons.

---

```
# 500 echantillons de taille 10
> set.seed(12) # repetabilite
> n=20
> m=500

> quotients <- c()
> for(k in 1:m){
  ech <- gapminder.QRD[sample(1:N,n,
    replace=FALSE),c("life_expectancy","lgdppc")]
  quotients[k] <- mean(ech$life_expectancy/ech$lgdppc)
}
```

---

La moyenne des 500 estimateurs est  $\bar{r} = 9.23$ :

---

```
> quotients <- data.frame(quotients)
> mean(quotients)
```

---

```
[1] 9.238648
```

On sait que  $\mu_X = 7.84$ . Il serait raisonnable de s'attendre à ce que  $\mu_Y \approx \bar{r}\mu_X = 72.431$ .

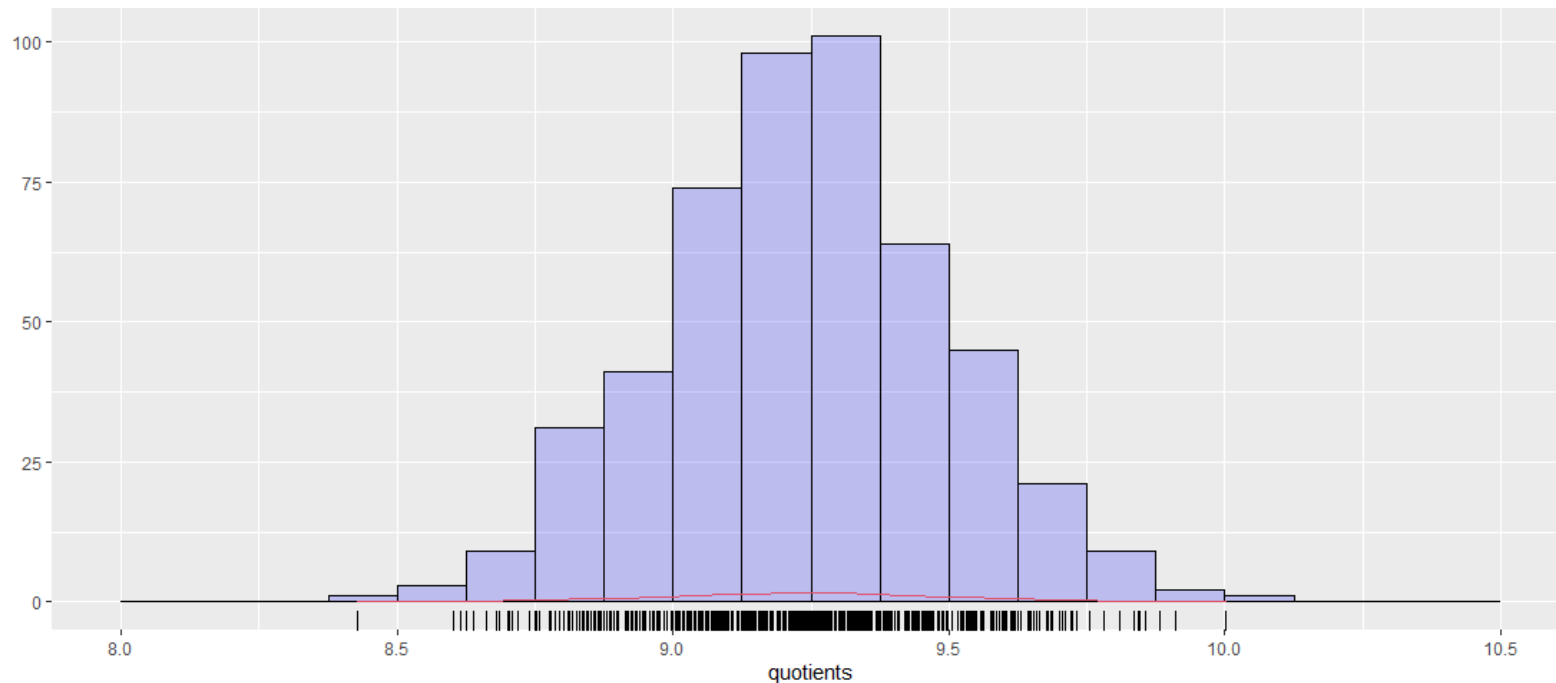
Est-ce une meilleure approximation que celle que nous avons obtenue au début du chapitre:  $\mu_Y \approx 68.00$ ?

C'est une question à laquelle on ne peut répondre sans connaître la distribution de l'estimateur  $r$ .

---

```
> ggplot(quotients, aes(quotients)) + geom_rug(aes(quotients)) +  
  geom_histogram(breaks=seq(8, 10.5, by = .125), col="black",  
    fill="blue", alpha=.2) + geom_density(col=2)
```

---



### 4.2.4 – I.C. de l'estimateur du quotient

On peut montrer que l'estimateur  $r$  suit **approximativement** une loi normale  $\mathcal{N}(E(r), V(r))$ , d'où la **marge d'erreur sur l'estimation** est

$$B_R \approx \hat{B}_R = 2\sqrt{\hat{V}(r)} \approx 2\sqrt{\frac{1}{\mu_X^2} \cdot \frac{s_W^2}{n} \left(1 - \frac{n}{N}\right)} \approx 2\sqrt{\frac{1}{\bar{x}^2} \cdot \frac{s_W^2}{n} \left(1 - \frac{n}{N}\right)},$$

où

$$s_W^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2,$$

d'où

$$\text{IC}(R; 0.95) : \quad r \pm \hat{B}_R$$

forme un **intervalle de confiance de  $R$  à environ 95%**.



On remarque que

$$\begin{aligned}\sigma_W^2 &= \frac{1}{N} \sum_{j=1}^N W_j^2 = \frac{1}{N} \sum_{j=1}^N (Y_j - RX_j)^2 = \frac{1}{N} \sum_{j=1}^N (Y_j - \mu_Y + \mu_Y - RX_j) \\ &= \frac{1}{N} \sum_{j=1}^N (Y_j - \mu_Y + R\mu_X - RX_j)^2 = \frac{1}{N} \sum_{j=1}^N [(Y_j - \mu_Y) - R(X_j - \mu_X)]^2 \\ &= \frac{1}{N} \sum_{j=1}^N (Y_j - \mu_Y)^2 - 2R \frac{1}{N} \sum_{j=1}^N (X_j - \mu_X)(Y_j - \mu_Y) + R^2 \frac{1}{N} \sum_{j=1}^N (X_j - \mu_X)^2 \\ &= \sigma_Y^2 - 2RCov(X, Y) + R^2 \sigma_X^2 = \sigma_Y^2 - 2R\rho\sigma_X\sigma_Y + R^2\sigma_X^2,\end{aligned}$$

$$\text{où } \rho = \frac{Cov(X, Y)}{\sigma_X\sigma_Y}.$$

Par analogie, nous avons alors  $s_W^2 = s_Y^2 - 2r\hat{\rho}s_Xs_Y + r^2s_X^2$ , où

$$s_X^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right), \quad s_Y^2 = \frac{1}{n-1} \left( \sum_{i=1}^n y_i^2 - n\bar{y}^2 \right),$$
$$s_{XY} = \frac{1}{n-1} \left( \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \right), \quad \text{et } \hat{\rho} = \frac{s_{XY}}{s_X s_Y}.$$

En pratique, on peut aussi utiliser la formule suivante:

$$s_W^2 = \frac{1}{n-1} \left( \sum_{i=1}^n y_i^2 - 2r \sum_{i=1}^n x_i y_i + r^2 \sum_{i=1}^n x_i^2 \right).$$

**Exemple:**

Soit un EAS  $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  de taille  $n = 132$  prélevé d'une population de taille  $N = 37,444$ . Donner un I.C. de  $R$  à environ 95% si

$$\sum_{i=1}^n x_i = 9464.6, \quad \sum_{i=1}^n y_i = 14691.6,$$
$$\sum_{i=1}^n x_i^2 = 686773.2, \quad \sum_{i=1}^n x_i y_i = 1062186, \quad \sum_{i=1}^n y_i^2 = 1670194.$$

**Solution:** Dans cet échantillon, nous obtenons  $r = \frac{14691.6}{9464.6} \approx 1.55$ , d'où

$$s_W^2 = \frac{1670194 - 2(1.55)(1062186) + (1.55)^2(686773.2)}{132 - 1} \approx 209.2 \text{ et}$$

$$\hat{V}(r) \approx \frac{132^2}{9464.6^2} \frac{209.2}{132} \left(1 - \frac{132}{37444}\right) = 0.0003 \implies \text{IC}(R; 0.95) \approx 1.552 \pm 0.035.$$

**Exemple:**

Donner un intervalle de confiance à 95% du quotient de l'espérance de vie par le logarithme du produit national brut en 2011 en utilisant un EAS de taille  $n = 20$ .

**Solution:** On commence par préparer l'ensemble de données.

---

```
> gapminder.QRD <- gapminder %>% filter(year==2011) %>%  
  select(life_expectancy,gdp,population)  
> gapminder.QRD <- gapminder.QRD[complete.cases(gapminder.QRD),]  
> N=nrow(gapminder.QRD)  
> gapminder.QRD <- gapminder.QRD %>% mutate(lgdppc=log(gdp/population))  
> (R = mean(gapminder.QRD$life_expectancy)/mean(gapminder.QRD$lgdppc))
```

---

```
[1] 9.046742
```

On prépare ensuite un échantillon de taille  $n = 20$ , et on calcule les sommes intermédiaires:

---

```
> n=20
> set.seed(123456) # repetabilite
> index = sample(1:N,n, replace=FALSE)
> ech = gapminder.QRD[index,c("life_expectancy","lgdppc")]

> (somme.xi = sum(ech$lgdppc))
> (somme.yi = sum(ech$life_expectancy))
> (somme.xi.2 = sum(ech$lgdppc^2))
> (somme.yi.2 = sum(ech$life_expectancy^2))
> (somme.xiyi = sum(ech$lgdppc*ech$life_expectancy))
```

---

```
[1] 167.2794 1450.82 1430.912 106117.4 12245.93
```

Finalement, on calcule l'estimateur  $r$  et sa variance, ainsi que l'intervalle de confiance recherché.

---

```
> r = somme.yi/somme.xi
> s2.W = 1/(n-1)*(somme.yi.2-2*r*somme.xiyi+r^2*somme.xi.2)
> V = n^2/somme.xi^2*(1/n)*s2.W*(1-n/N)
> B = 2*sqrt(V)
> c(r-B,r+B)
```

---

```
[1] 8.252515 9.093552
```

On s'attend alors à ce que le quotient  $R$  se retrouve dans  $(8.25, 9.09)$  avec 95% de probabilité: puisque  $R = 9.046742$ , c'est bien le cas.

Comme nous l'avons remarqué à plusieurs reprises, l'intervalle de confiance peut changer en fonction de l'échantillon prélevé.

## 4.2.5 – Estimation de la moyenne et du total

En pratique, on connaît souvent  $\tau_X$  et/ou  $\mu_X$ . Il est possible de se servir de la relation

$$\mu_Y = R\mu_X, \quad \text{où } R = \frac{\mu_Y}{\mu_X}$$

afin d'approximer  $\mu_Y$  (si  $\mu_X$  est inconnue, on utilise  $\mu_X \approx \bar{x}$ ).

Puisque  $r = \bar{y}/\bar{x}$ , nous obtenons l'estimateur  $\hat{\mu}_{Y;R}$  par le quotient:

$$\hat{\mu}_{Y;R} = r \cdot \mu_X.$$

Mais nous avons déjà observé que  $r$  est un estimateur **biaisé** de  $R$ , c'est donc dire que l'on s'attend à ce que  $\hat{\mu}_{Y;R}$  soit un estimateur **biaisé** de  $\mu_Y$ , suivant une loi normale:  $\hat{\mu}_{Y;R} \sim_{\text{approx}} \mathcal{N}(E(\hat{\mu}_{Y;R}), V(\hat{\mu}_{Y;R}))$ .

On calcule aisément que

$$\mathbb{E}[\hat{\mu}_{Y;R} - \mu_Y] = \mu_X \mathbb{E}[r - R] \approx \frac{1}{\mu_X} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n} \left( \frac{N-n}{N-1} \right)$$

$$V(\hat{\mu}_{Y;R}) = V(r \cdot \mu_X) = \mu_X^2 V(r) \approx \frac{\sigma_W^2}{n} \left( \frac{N-n}{N-1} \right).$$

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

$$B_{\mu_{Y;R}} \approx \hat{B}_{\mu_{Y;R}} = 2\sqrt{V(\hat{\mu}_{Y;R})} \approx 2\sqrt{\frac{s_W^2}{n} \left( 1 - \frac{n}{N} \right)}, \quad s_W^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2,$$

d'où  $IC_R(\mu_Y; 0.95) \equiv \hat{\mu}_{Y;R} \pm \hat{B}_{\mu_{Y;R}}$  forme un **intervalle de confiance de  $\mu_Y$  à environ 95%**.



Il est également possible de se servir de la relation

$$\tau_Y = R\tau_X, \quad \text{où } R = \frac{\mu_Y}{\mu_X} = \frac{\tau_Y}{\tau_X}$$

afin d'approximer  $\tau_Y$  (si  $\tau_X$  est inconnu, on utilise  $\tau_X \approx N\bar{x}$ ).

Puisque  $r = \bar{y}/\bar{x}$ , nous obtenons l'**estimateur**  $\hat{\tau}_{Y;R}$  **par le quotient**:

$$\hat{\tau}_{Y;R} = r \cdot \tau_X.$$

Mais nous avons déjà observé que  $r$  est un estimateur **biasé** de  $R$ , c'est donc dire que l'on s'attend à ce que  $\hat{\tau}_{Y;R}$  soit un estimateur **biaisé** de  $\tau_Y$ , suivant une loi normale:

$$\hat{\tau}_{Y;R} \sim_{\text{approx}} \mathcal{N}(\mathbb{E}(\hat{\tau}_{Y;R}), \mathbb{V}(\hat{\tau}_{Y;R})).$$

On calcule aisément que

$$\mathbb{E}[\hat{\tau}_{Y;R} - \tau_Y] = \tau_X \mathbb{E}[r - R] \approx \frac{N}{\mu_X} \cdot \frac{R\sigma_X^2 - \rho\sigma_X\sigma_Y}{n} \left( \frac{N-n}{N-1} \right)$$

$$V(\hat{\tau}_{Y;R}) = V(r \cdot \tau_X) = \tau_X^2 V(r) = N^2 \mu_X^2 V(r) \approx N^2 \cdot \frac{\sigma_W^2}{n} \left( \frac{N-n}{N-1} \right).$$

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

$$B_{\tau_{Y;R}} \approx \hat{B}_{\tau_{Y;R}} = 2\sqrt{V(\hat{\tau}_{Y;R})} \approx 2N\sqrt{\frac{s_W^2}{n} \left( 1 - \frac{n}{N} \right)}, \quad s_W^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - rx_i)^2,$$

d'où  $\text{IC}_R(\tau_Y; 0.95) \equiv \hat{\tau}_{Y;R} \pm \hat{B}_{\tau_{Y;R}}$  forme un **intervalle de confiance de  $\tau_Y$  à environ 95%**.

**Exemple:**

Soit un EAS  $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  de taille  $n = 132$  prélevé d'une population de taille  $N = 37,444$ . Donner un I.C. de  $\mu_Y$  à environ 95% par le quotient si

$$\sum_{i=1}^n x_i = 9464.6, \quad \sum_{i=1}^n y_i = 14691.6,$$
$$\sum_{i=1}^n x_i^2 = 686773.2, \quad \sum_{i=1}^n x_i y_i = 1062186, \quad \sum_{i=1}^n y_i^2 = 1670194.$$

**Solution:** Nous avons déjà calculé  $r \approx 1.55$ ,  $s_W^2 \approx 209.2$ ,  $\hat{V}(r) \approx 0.00031$  et  $\text{IC}(R; 0.95) \approx 1.552 \pm 0.035$ . De plus,  $\bar{x} = 9464.6/132 = 71.70$ . Ainsi,

$$\text{IC}_R(\mu_Y; 0.95) = \mu_X \cdot \text{IC}(R; 0.95) \approx \bar{x} \cdot \text{IC}(R; 0.95) \equiv 111.29 \pm 2.51.$$

**Exemple:**

Donner un intervalle de confiance à 95% de l'espérance de vie moyenne par pays en 2011 en utilisant un EAS de taille  $n = 20$  et la méthode du quotient par le logarithme du PNB ( $X$ ).

**Solution:** Nous utilisons le même échantillon qu'à l'exemple aux pp. 27-29. Nous avons déjà obtenu un intervalle de confiance pour le quotient:

$$IC(R; 0.95) = (8.25, 9.09).$$

La moyenne empirique pour l'échantillon était  $\bar{x} = \frac{167.2794}{20} = 8.364$ . L'intervalle de confiance pour l'espérance de vie moyenne est ainsi

$$IC_R(\mu_Y; 0.95) = \mu_X \cdot IC(R; 0.95) \approx \bar{x} \cdot (8.25, 9.09) = (69.00, 76.03).$$

La valeur réelle est  $\mu_Y = 70.95$ .

## 4.2.6 – Taille de l'échantillon

Lorsque l'on cherche à estimer  $R$ , nous obtenons

$$\begin{aligned}
 B_R &\approx 2\sqrt{\frac{1}{\mu_X^2} \cdot \frac{\sigma_W^2}{n} \left(\frac{N-n}{N-1}\right)} \iff \underbrace{\frac{B_R^2 \mu_X^2}{4}}_{=D_R} = \frac{\sigma_W^2}{n} \left(\frac{N-n}{N-1}\right) \\
 &\iff \frac{(N-1)D_R}{\sigma_W^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\
 &\iff \frac{(N-1)D_R + \sigma_W^2}{\sigma_W^2} = \frac{N}{n} \\
 &\iff n_R = \frac{N\sigma_W^2}{(N-1)D_R + \sigma_W^2}.
 \end{aligned}$$

Si on cherche à approximer  $\mu_Y$ , nous obtenons de même:

$$B_{\mu_Y;R} \approx 2\sqrt{\frac{\sigma_W^2}{n} \left( \frac{N-n}{N-1} \right)} \iff n_{\mu_Y} = \frac{N\sigma_W^2}{(N-1)D_{\mu_Y} + \sigma_W^2}, \quad D_{\mu_Y} = \frac{B_{\mu_Y;R}^2}{4},$$

et si on cherche à approximer  $\tau_Y$ , nous obtenons de même:

$$B_{\tau_Y;R} \approx 2\sqrt{N^2 \cdot \frac{\sigma_W^2}{n} \left( \frac{N-n}{N-1} \right)} \iff n_{\tau_Y} = \frac{N\sigma_W^2}{(N-1)D_{\tau_Y} + \sigma_W^2}, \quad D_{\tau_Y} = \frac{B_{\tau_Y;R}^2}{4N^2}.$$

Puisque nous ne connaissons pas  $\sigma_W^2$  en général, on prélève souvent un échantillon préliminaire de petite taille et on utilise la variance empirique  $s_W^2$  en tant qu'estimateur de  $\sigma_W^2$ .

**Exemple:**

Soit un EAS  $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  de taille  $n$ , prélevé d'une population de taille  $N = 37,444$ .

Lors d'une étude préalable, nous avons déterminé que  $\sigma_W^2 \approx 209.2$  et  $\mu_X \approx 71.7$ .

Déterminer la taille minimale de l'échantillon requise afin de s'assurer que la marge d'erreur sur l'estimation

1. du quotient  $R$  par  $r$  soit au plus 0.025;
2. de la moyenne  $\mu_Y$  par  $\hat{\mu}_{Y;R}$  soit au plus 5, et
3. du total  $\tau_Y$  par  $\hat{\tau}_{Y;R}$  soit au plus 25.

**Solution:** Nous nous servons simplement des formules.

1. Puisque  $D_R = \frac{B_R^2 \mu_X^2}{4} = \frac{0.025^2 (71.7)^2}{4} \approx 0.8033$ , alors

$$n_R = \frac{37444(209.2)}{(37444 - 1)(0.8033) + 209.2} = 258.6453 \implies n_R \geq 259.$$

2. Puisque  $D_{\mu_Y} = \frac{B_{\mu_Y;R}^2}{4} = \frac{5^2}{4} \approx 6.25$ , alors

$$n_{\mu_Y} = \frac{37444(209.2)}{(37444 - 1)(6.25) + 209.2} = 33.443 \implies n_{\mu_Y} \geq 34.$$



3. Puisque  $D_{\tau_Y} = \frac{B_{\tau_Y;R}^2}{4N^2} = \frac{25^2}{4(37444)} \approx 0.001502243$ , alors

$$n_{\tau_Y} = \frac{37444(209.2)}{(37444 - 1)(0.001502243) + 209.2} = 29509.62 \implies n_{\tau_Y} \geq 29510.$$

La marge d'erreur recherchée  $B_{\tau_Y;R}$  est sans doute trop serrée.

### Exemple:

Déterminer la taille  $n$  de l'échantillon afin d'estimer l'espérance de vie moyenne  $\mu_Y$  à l'aide du quotient  $R$  par le log du PNB en 2011, avec une marge sur l'erreur d'estimation  $B_{\mu_Y;R} = 1$ , si  $\sigma_W^2 \approx 70.2$  et  $N = 168$ .

**Solution:** Selon la formule, nous obtenons:

$$n_{\mu_Y} = \frac{168(70.2)}{167(1^2/4) + 70.2} = 105.35 \implies n_{\mu_Y} \geq 106.$$

## 4.3 – Estimation par la régression

L'estimation par le quotient est un cas spécial d'une méthode plus générale, l'estimation par la régression.

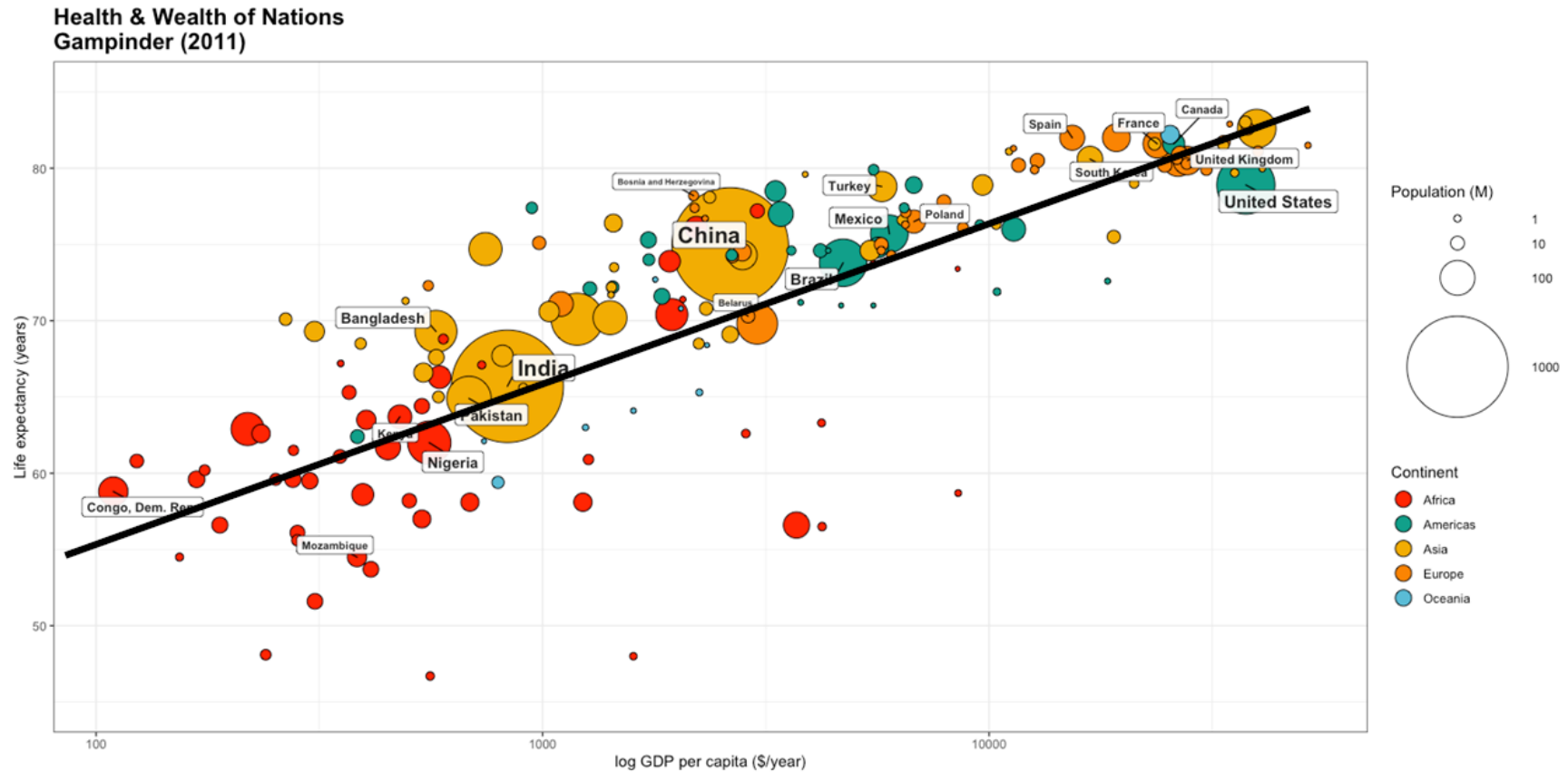
Dans l'ensemble de données `gapminder.csv` pour 2011, on reconnaît l'existence d'une relation plus ou moins linéaire entre l'**espérance de vie**  $Y$  et le **logarithme du PNB par habitant**  $X$  pour  $N = 168$  pays.

Lorsque l'on calcule

$$r = \bar{y}/\bar{x}$$

à l'aide d'un EAS de taille  $n$ , on suppose que la relation réelle entre  $Y$  et  $X$  prend la forme  $Y = RX \approx rX$ , c'est-à-dire que c'est une droite de pente  $r$  **passant par l'origine**.

Cette dernière condition ne semble pas être remplie. Que faire dans ce cas?



### 4.3.1 – Estimateur par la régression

Soient

- $\mathcal{U} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$  une population finie de taille  $N$  pour laquelle chaque unité  $u_j$  admet 2 réponses ( $X_j$  et  $Y_j$ ), et
- $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\} \subseteq \mathcal{U}$  un **échantillon aléatoire simple bivarié** de taille  $n$ .

On suppose que la relation entre  $Y$  et  $X$  prend la forme

$$Y - \mu_Y = \beta(X - \mu_X).$$

Si  $\mu_X$  est connue (comme nous l'avons supposé lors de l'estimation par le quotient), l'**estimateur par la régression**  $\hat{\mu}_{Y;L}$  de  $\mu_Y$  donné par l'EAS  $\mathcal{Y}$  prend la forme

$$\hat{\mu}_{Y;L} = \bar{y} + \beta(\mu_X - \bar{x}).$$

Pour l'instant,  $\beta$  est une constante **inconnue** (puisque  $\mu_Y$  est inconnue). Avec cette hypothèse, les caractéristiques de la distribution de l'estimateur par la régression  $\hat{\mu}_{Y;L}$  sont les suivantes.

Puisque  $\mathcal{Y}$  est un EAS,  $E(\bar{x}) = \mu_X$  et  $E(\bar{y}) = \mu_Y$ , d'où

$$E(\hat{\mu}_{Y;L}) = E(\bar{y}) + \beta(\mu_X - E(\bar{x})) = \mu_Y + \beta(\mu_X - \mu_X) = \mu_Y.$$

Considérons la variable aléatoire  $W = Y + \beta(\mu_X - X)$ . Nous avons alors

$$\mu_W = \mu_Y + \beta(\mu_X - \mu_X) = \mu_Y.$$

La moyenne de  $W$  dans l'échantillon EAS  $\mathcal{Y}$  est alors

$$\bar{w} = \bar{y} + \beta(\mu_X - \bar{x}) = \hat{\mu}_{Y;L} \implies V(\hat{\mu}_{Y;L}) = V(\bar{w}) = \frac{\sigma_{W;L}^2}{n} \left( \frac{N-n}{N-1} \right).$$

Mais

$$\begin{aligned} \sigma_{W;L}^2 &= \frac{1}{N} \sum_{j=1}^N (W_j - \mu_W)^2 = \frac{1}{N} \sum_{j=1}^N (Y_j + \beta(\mu_X - X_j) - \mu_Y)^2 \\ &= \frac{1}{N} \sum_{j=1}^N \{(Y_j - \mu_Y) - \beta(X_j - \mu_X)\}^2 = \sigma_Y^2 - 2\beta\rho\sigma_X\sigma_Y + \beta^2\sigma_X^2, \end{aligned}$$

$$\text{où } \rho = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Ainsi,

$$V(\hat{\mu}_{Y;L}) = \frac{\sigma_Y^2 - 2\beta\rho\sigma_X\sigma_Y + \beta^2\sigma_X^2}{n} \left( \frac{N-n}{N-1} \right).$$

En général, étant donné une erreur systématique (biais) donnée, la préférence va à l'estimateur **possédant la plus faible variance**. La valeur de  $\beta$  qui minimise  $V(\hat{\mu}_{Y;L})$  satisfait donc à

$$\frac{\partial V(\hat{\mu}_{Y;L})}{\partial \beta}(\beta^*) = \frac{1}{n} \left( \frac{N-n}{N-1} \right) (-2\rho\sigma_X\sigma_Y + 2\beta^*\sigma_X^2) = 0,$$

c'est-à-dire que

$$\beta^* = \rho \frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y} \cdot \frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2},$$

d'où

$$\begin{aligned} V(\hat{\mu}_{Y;L}) &= \frac{\sigma_Y^2 - 2\beta^* \rho \sigma_X \sigma_Y + (\beta^*)^2 \sigma_X^2}{n} \left( \frac{N-n}{N-1} \right) \\ &= \frac{\sigma_Y^2 - 2\rho \frac{\sigma_Y}{\sigma_X} \rho \sigma_X \sigma_Y + (\rho \frac{\sigma_Y}{\sigma_X})^2 \sigma_X^2}{n} \left( \frac{N-n}{N-1} \right) \\ &= \frac{\sigma_Y^2 - 2\rho^2 \sigma_Y^2 + \rho^2 \sigma_Y^2}{n} \left( \frac{N-n}{N-1} \right) \\ &= \frac{\sigma_Y^2 (1 - \rho^2)}{n} \left( \frac{N-n}{N-1} \right). \end{aligned}$$

Si  $\sigma_Y$  et  $\rho$  sont connues, l'expression est exacte. Autrement, ce n'est pas sans rappeler le contexte de la **régression linéaire simple**.



### 4.3.2 – Biais de l'estimateur par la régression

Étant donnés  $n$  points  $(x_i, y_i)$ , il s'agit de déterminer les constantes  $\alpha$  et  $\beta$  qui expriment le mieux une relation linéaire entre  $X$  et  $Y$

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

où l'on suppose que  $(\varepsilon_1, \dots, \varepsilon_n) \sim_{\text{approx.}} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ . Il y a plusieurs manières d'interpréter la phrase “le mieux” – les **estimateurs de moindres carrés**  $\hat{\alpha}$  et  $\hat{\beta}$  sont ceux qui minimisent la somme des carrés des résidus

$$Q(\alpha, \beta) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2.$$

On résoud le système d'équations

$$\frac{\partial Q}{\partial \alpha}(a, b) = \sum_{i=1}^n -2(y_i - a - bx_i) = 0, \quad \frac{\partial Q}{\partial \beta}(a, b) = \sum_{i=1}^n -2x_i(y_i - a - bx_i) = 0$$

afin d'obtenir

$$\hat{\alpha} = a = \bar{y} - b\bar{x} \quad \text{et} \quad \hat{\beta} = b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

En pratique, c'est ce  $b = \hat{\rho} \frac{s_Y}{s_X}$  qui joue le rôle de l'estimateur de  $\beta^*$ . On note qu'il varie d'un EAS à l'autre.

Puisque  $b$  n'est pas constant, nous ne pouvons conclure que

$$E(b\bar{x}) = E(b) E(\bar{x}),$$

d'où

$$E(\hat{\mu}_{Y;L}) = E(\bar{y}) + \mu_X E(b) - E(b\bar{x}) \neq \mu_Y,$$

en général.

Cependant, si la taille  $n$  de l'échantillon est élevée, il est possible de montrer que

$$E[\hat{\mu}_{Y;L} - \mu_Y]$$

est d'ordre  $\frac{1}{n}$  (comme c'était le cas pour l'erreur systématique de l'estimateur par le quotient):  $\hat{\mu}_{Y;L}$  est donc un **estimateur biaisé** de  $\mu_Y$ .

### 4.3.3 – Variabilité de l'estimateur par la régression

La **variance de l'estimateur**  $\hat{\mu}_{Y;L}$  est aussi d'ordre  $\frac{1}{n}$ , d'où le rapport entre l'erreur systématique  $E[\hat{\mu}_{Y;L} - \mu_Y]$  et l'écart type de  $\hat{\mu}_{Y;L}$  est d'ordre  $\frac{1}{\sqrt{n}}$ .

Ainsi, lorsque  $n, N \rightarrow \infty$  (en supposant que  $N \gg n$ ), nous obtenons

$$\frac{E[\hat{\mu}_{Y;L} - \mu_Y]}{ET(\hat{\mu}_{Y;L})} \rightarrow 0.$$

Quoiqu'il soit impossible de se débarrasser du biais, l'erreur d'estimation

$$EQM(\hat{\mu}_{Y;L}) = V(\hat{\mu}_{Y;L}) + (E(\hat{\mu}_{Y;L}) - \mu_Y)^2$$

est donc dominée par la variance  $V(\hat{\mu}_{Y;L})$  si  $n$  est **suffisamment élevée**.

### 4.3.4 – I.C. de l'estimateur par la régression

L'estimateur  $\hat{\mu}_{Y;L}$  suit **approximativement** une loi normale  $\mathcal{N}(E(\hat{\mu}_{Y;L}), V(\hat{\mu}_{Y;L}))$ , d'où la **marge d'erreur sur l'estimation** est

$$B_L \approx \hat{B}_L = 2\sqrt{\hat{V}(\hat{\mu}_{Y;L})} \approx 2\sqrt{\frac{s_{W;L}^2}{n} \left(1 - \frac{n}{N}\right)},$$

où  $s_{W;L}^2$  est l'**erreur quadratique moyenne de la régression linéaire**,

$$s_{W;L}^2 = \frac{n-1}{n-2}(s_Y^2 - b^2 s_X^2) = \frac{n-1}{n-2} \cdot s_Y^2(1 - \hat{\rho}^2),$$

d'où  $IC_L(\mu_Y; 0.95) : \hat{\mu}_{Y;L} \pm \hat{B}_L$  forme un **intervalle de confiance de  $\mu_Y$  à environ 95%** (on passe à  $\tau_Y$  et  $p_Y$  de la manière habituelle).

**Exemple:**

Soit un EAS  $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  de taille  $n = 132$  prélevé d'une population de taille  $N = 37,444$ .

Lors d'une étude préalable, il est déterminé que  $\mu_X \approx 70.3$ .

Donner un I.C. de  $\mu_Y$  à environ 95% par la régression si

$$\sum_{i=1}^n x_i = 9464.6, \quad \sum_{i=1}^n y_i = 14691.6,$$
$$\sum_{i=1}^n x_i^2 = 686773.2, \quad \sum_{i=1}^n x_i y_i = 1062186, \quad \sum_{i=1}^n y_i^2 = 1670194.$$

**Solution:** Nous devons calculer  $\bar{x}$ ,  $\bar{y}$ ,  $s_X^2$ ,  $s_{XY}$ ,  $s_Y^2$ , et  $\hat{\rho}$ . Mais

$$\bar{x} = \frac{9464.6}{132} \approx 71.7, \quad \bar{y} = \frac{14691.6}{132} \approx 111.3,$$

$$s_X^2 = \frac{686773.2 - 132(71.7)^2}{132 - 1} \approx 62.2, \quad s_Y^2 = \frac{1670194 - 132(111.3)^2}{132 - 1} \approx 267.3$$

$$s_{XY} = \frac{1062186 - 132(71.7)(111.3)}{132 - 1} \approx 67.2, \quad \hat{\rho} = \frac{67.2}{\sqrt{(62.2)(267.3)}} \approx 0.521.$$

L'estimateur de la pente de régression est donc  $b = \hat{\rho} \frac{s_Y}{s_X} = 1.08$ . De plus,

$$s_{W;L}^2 = \frac{131}{130} \cdot 267.3 \cdot (1 - 0.521^2) \approx 196.77.$$

C'est donc dire que

$$\hat{\mu}_{Y;L} = 111.3 + 1.08(\underbrace{70.3 - 71.7}_{\mu_X}) = 109.8$$

et

$$\hat{B}_L \approx 2\sqrt{\frac{196.77}{132} \left(1 - \frac{132}{37444}\right)} = 2.43,$$

d'où

$$\text{IC}_L(\mu_Y; 0.95) \equiv 109.8 \pm 2.43.$$

On se sert de la corrélation  $\hat{\rho}$  afin de réduire la taille de l'intervalle de confiance – on passe de  $\hat{B}_{\bar{y}} = 2.84$  (EAS) à  $\hat{B}_L = 2.43$ .

Bien sûr, cela dépend de la validité de la régression linéaire.



**Exemple:**

Donner un intervalle de confiance à 95% de l'espérance de vie moyenne par pays en 2011 en utilisant un EAS de taille  $n = 20$  et la méthode de la régression par rapport au logarithme du PNB (avec  $\mu_X = 7.84$ ).

**Solution:** On commence par préparer l'ensemble de données.

---

```
> gapminder.QRD <- gapminder %>% filter(year==2011) %>%  
  select(life_expectancy,gdp,population)  
> gapminder.QRD <- gapminder.QRD[complete.cases(gapminder.QRD),]  
> N=nrow(gapminder.QRD)  
> gapminder.QRD <- gapminder.QRD %>% mutate(lgdppc=log(gdp/population))  
> (mu.X = mean(gapminder.QRD[, "lgdppc"]))
```

---

```
[1] 7.842661
```

On prépare ensuite un échantillon de taille  $n = 20$ , et on calcule les quantités requises:

---

```
> n=20
> set.seed(123456)
> index = sample(1:N,n, replace=FALSE)
> ech = gapminder.QRD[index,c("life_expectancy","lgdppc")]

# moyennes d'échantillon
> (y.bar = mean(ech$life_expectancy))
> (x.bar = mean(ech$lgdppc))
```

---

```
[1] 72.331 7.952495
```

```
# sommes pour les variances
> somme.xi = sum(ech$lgdppc)
> somme.yi = sum(ech$life_expectancy)
> somme.xi.2 = sum(ech$lgdppc^2)
> somme.yi.2 = sum(ech$life_expectancy^2)
> somme.xiyi = sum(ech$lgdppc*ech$life_expectancy)

> s2.X = (somme.xi.2-n*x.bar^2)/(n-1)
> s2.Y = (somme.yi.2-n*y.bar^2)/(n-1)
> s.XY = (somme.xiyi-n*x.bar*y.bar)/(n-1)

# coefficient de correlation
> (rho = s.XY/sqrt(s2.X*s2.Y))
```

---

```
[1] 0.9398905
```

---

```
# EQM (MSE)
```

```
> (s2.W.L = (n-1)/(n-2)*s2.Y*(1-rho^2))
```

---

```
[1] 7.664038
```

La marge d'erreur est ainsi:

---

```
> (B = 2*sqrt(s2.W.L/n*(1-n/N)))
```

---

```
[1] 1.162037
```

et l'I.C. à 95% pour l'espérance de vie moyenne par pays est:

---

```
# estime ponctuel
```

```
> (hat.mu.Y.L = y.bar + rho*sqrt(s2.Y/s2.X)*(mu.X-x.bar))
```

---

```
[1] 71.8518
```

```
# limite inferieure  
> hat.mu.Y.L-B  
# limite superieure  
> hat.mu.Y.L+B
```

---

```
[1] 70.68976 73.01383
```

La valeur réelle est  $\mu_Y = 70.95$ .

On peut aussi calculer l'estimé et l'intervalle de confiance directement en utilisant la fonction `lm()`.

```
> reg.lin = lm(life_expectancy~lgdppc, data=ech)  
> summary(reg.lin)
```

---

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	37.6343	3.0350	12.40	2.97e-10	***
lgdppc	4.3630	0.3736	11.68	7.81e-10	***

Residual standard error: 2.768 on 18 degrees of freedom

Les valeurs recherchées peuvent être extraites comme suit:

---

```
> (b = as.numeric(reg.lin$coefficients[2]))
```

---

```
[1] 4.362993
```

---

```
> (s2.W.L = summary(reg.lin)$sigma^2)
```

---

```
[1] 7.664038
```

### 4.3.5 – Taille de l'échantillon

Lorsque l'on cherche à estimer  $\mu_Y$  par la régression, nous obtenons

$$\begin{aligned}
 B_L &\approx 2\sqrt{\frac{\sigma_{W;L}^2}{n} \left( \frac{N-n}{N-1} \right)} \iff \underbrace{\frac{B_L^2}{4}}_{=D_L} = \frac{\sigma_{W;L}^2}{n} \left( \frac{N-n}{N-1} \right) \\
 &\iff \frac{(N-1)D_L}{\sigma_{W;L}^2} = \frac{N-n}{n} = \frac{N}{n} - 1 \\
 &\iff \frac{(N-1)D_L + \sigma_{W;L}^2}{\sigma_{W;L}^2} = \frac{N}{n}, \\
 &\iff n_L = \frac{N\sigma_{W;L}^2}{(N-1)D_L + \sigma_{W;L}^2}.
 \end{aligned}$$

Si on cherche à approximer  $\tau_Y$ , nous obtenons de même:

$$B_{\tau;L} \approx 2N \sqrt{\frac{\sigma_{W;L}^2}{n} \left( \frac{N-n}{N-1} \right)} \iff n_{\tau;L} = \frac{N\sigma_{W;L}^2}{(N-1)D_{\tau;L} + \sigma_{W;L}^2},$$

où  $D_{\tau;L} = \frac{B_{\tau;L}^2}{4N^2}.$

Puisque nous ne connaissons pas  $\sigma_{W;L}^2$  en général, on prélève souvent un échantillon préliminaire de petite taille et on utilise l'erreur quadratique moyenne de la régression  $s_{W;L}^2$  en tant qu'estimateur de  $\sigma_{W;L}^2$ .

**⚠** Même si les manipulations formelles peuvent toujours s'effectuer, **l'estimé risque de ne pas être valide si la relation entre les variables  $X$  et  $Y$  n'est pas linéaire.**



**Exemple:**

Soit un EAS  $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  de taille  $n$ , prélevé d'une population de taille  $N = 37,444$ .

Lors d'une étude préalable, nous avons déterminé que  $\sigma_{W;L}^2 \approx 188.2$ .

Déterminer la taille minimale de l'échantillon requise afin de s'assurer que la marge d'erreur sur l'estimation par la régression

1. de la moyenne  $\mu_Y$  par  $\hat{\mu}_{Y;L}$  soit au plus 5, et
2. du total  $\tau_Y$  par  $\hat{\tau}_{Y;L}$  soit au plus 250.

**Solution:** Nous nous servons simplement des formules.

1. Puisque  $D_L = \frac{B_L^2}{4} = \frac{5^2}{4} \approx 6.25$ , alors

$$n_L = \frac{37444(188.2)}{(37444 - 1)(6.25) + 188.2} = 28.39 \implies n_L \geq 29.$$

2. Puisque  $D_{\tau_Y;L} = \frac{B_{\tau_Y;L}^2}{4N^2} = \frac{250^2}{4(37444)} \approx 0.4172898$ , alors

$$n_{\tau_Y;L} = \frac{37444(188.2)}{(37444 - 1)(0.4172898) + 188.2} = 445.6497 \implies n_{\tau_Y} \geq 446.$$

En supposant, bien sûr que l'hypothèse de linéarité soit satisfaite.

**Exemple:**

Déterminer la taille  $n$  de l'échantillon afin d'estimer l'espérance de vie moyenne  $\mu_Y$  à l'aide de la régression par le log du PNB en 2011, avec une marge sur l'erreur d'estimation  $B_L = 1$ , si  $\sigma_{W;L} \approx 5.194$  et  $N = 168$ .

**Solution:** Selon la formule, nous obtenons:

$$n_L = \frac{168(5.194)^2}{167(1^2/4) + (5.194)^2} = 65.94498 \implies n_L \geq 66.$$

Puisque il y a de bonnes raison de croire que la relation entre l'espérance de vie et le log du PNB en 2011 est approximativement linéaire (cf. graphique), l'approche par la régression est recommandée (en supposant, bien sûr, que  $\mu_X$  soit connue) – comparer avec l'exemple qui utilise la méthode du quotient.

## 4.4 – Estimation par la différence

L'estimation **par la différence** est un autre cas spécial de l'estimation par la régression, où l'on suppose que la pente  $\beta$  prend la valeur 1.

Si  $\mu_X$  est connue, l'**estimateur par la différence**  $\hat{\mu}_{Y;D}$  de  $\mu_Y$  donné par l'EAS  $\mathcal{Y}$  prend la forme

$$\hat{\mu}_{Y;D} = \bar{y} + (\mu_X - \bar{x}).$$

L'estimation par la différence est une bonne stratégie lorsque la relation entre  $X$  et  $Y$  est à peu près **linéaire** et de **pente 1** (passant ou non **par l'origine**), et tant que la variance de  $Y$  **le long de cette droite est constante pour tout  $X$** .

Avec cette hypothèse, les caractéristiques de la distribution de l'estimateur par la différence  $\hat{\mu}_{Y;D}$  sont les suivantes.

Puisque  $\mathcal{Y}$  est un EAS,  $E(\bar{x}) = \mu_X$  et  $E(\bar{y}) = \mu_Y$ , d'où

$$E(\hat{\mu}_{Y;D}) = E(\bar{y}) + (\mu_X - E(\bar{x})) = \mu_Y + (\mu_X - \mu_X) = \mu_Y.$$

Considérons la variable aléatoire  $D = Y - X$ , dont l'espérance est

$$\mu_D = \mu_Y - \mu_X.$$

La moyenne empirique de  $D$  dans l'échantillon EAS  $\mathcal{Y}$  est alors

$$\bar{d} = \bar{y} - \bar{x} \implies \hat{\mu}_{Y;D} = \mu_X + (\bar{y} - \bar{x}) = \mu_X + \bar{d}.$$

Par conséquent,

$$V(\hat{\mu}_{Y;D}) = V(\mu_X + \bar{d}) = V(\bar{d}) = \frac{\sigma_D^2}{n} \left( \frac{N-n}{N-1} \right).$$

Mais

$$\begin{aligned} \sigma_D^2 &= \frac{1}{N} \sum_{j=1}^N (D_j - \mu_D)^2 = \frac{1}{N} \sum_{j=1}^N \{(Y_j - X_j) - (\mu_Y - \mu_X)\}^2 \\ &= \frac{1}{N} \sum_{j=1}^N \{(Y_j - \mu_Y) - (X_j - \mu_X)\}^2 = \sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2, \end{aligned}$$

$$\text{où } \rho = \frac{\text{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Ainsi,

$$V(\hat{\mu}_{Y;D}) = \frac{\sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2}{n} \left( \frac{N-n}{N-1} \right).$$

L'estimateur  $\hat{\mu}_{Y;D}$  suit **approximativement** une loi normale  $\mathcal{N}(E(\hat{\mu}_{Y;D}), V(\hat{\mu}_{Y;D}))$ , d'où la **marge d'erreur sur l'estimation** est

$$B_D \approx \hat{B}_D = 2\sqrt{\hat{V}(\hat{\mu}_{Y;D})} \approx 2\sqrt{\frac{s_D^2}{n} \left( 1 - \frac{n}{N} \right)},$$

où

$$s_D^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2 = s_Y^2 - 2\hat{\rho}s_Xs_Y + s_X^2,$$

d'où  $IC_D(\mu_Y; 0.95) : \hat{\mu}_{Y;D} \pm \hat{B}_D$  forme un **intervalle de confiance de  $\mu_Y$  à environ 95%** (on passe à  $\tau_Y$  et  $p_Y$  de la manière habituelle).

**Exemple:**

Les auditeurs s'intéressent souvent à la comparaison de la valeur vérifiée  $Y$  des articles avec leur valeur comptable  $X$ . Supposons que  $N = 180$  articles en stock aient une valeur comptable de  $\tau_X = 13,320\$$ . Un EAS de  $n = 10$  articles permet d'obtenir les données suivantes.

<b>Article</b> $i$	1	2	3	4	5	6	7	8	9	10
<b>Vérifiée</b> $y_i$	9	14	7	29	45	109	40	238	60	170
<b>Comptable</b> $x_i$	10	12	8	26	47	112	36	240	59	167
$d_i = y_i - x_i$	-1	2	-1	3	-2	-3	4	-2	1	3

Donner un I.C. de la valeur moyenne des vérifications  $\mu_Y$  à environ 95% par la différence.



**Solution:** Nous devons calculer  $\bar{d}$  et  $s_D^2$ . Mais

$$\sum_{i=1}^{10} d_i = 4, \quad \sum_{i=1}^{10} d_i^2 = 58, \implies \bar{d} = \frac{4}{10} \quad \text{et} \quad s_D^2 = \frac{58 - 10(0.4)^2}{10 - 1} = 6.27.$$

Puisque  $\mu_X = \frac{\tau_X}{N} = \frac{13320}{180} = 74$ , l'estimateur de la valeur moyenne des vérifications est

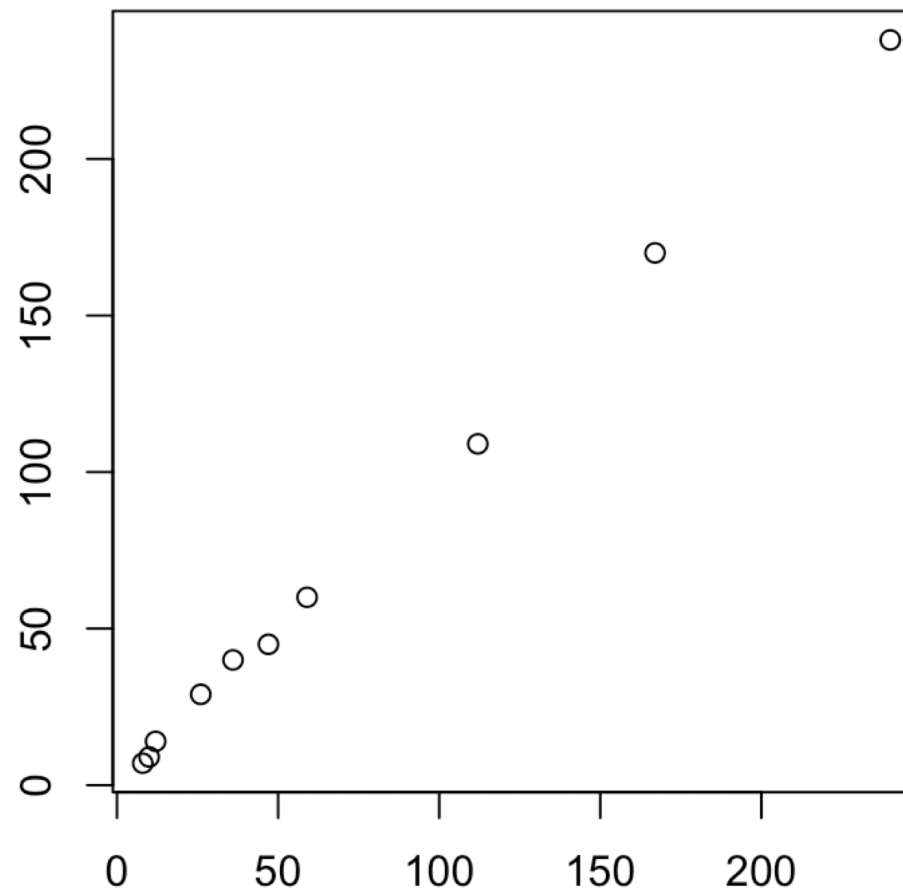
$$\hat{\mu}_{Y;D} = \mu_X + \bar{d} = 74 + 0.4 = 74.4$$

et

$$\hat{B}_D \approx 2\sqrt{\hat{V}(\hat{\mu}_D)} = 2\sqrt{\frac{6.27}{10} \left(1 - \frac{10}{180}\right)} = 1.54,$$

d'où

$$\text{IC}_D(\mu_Y; 0.95) : \quad 74.4 \pm 1.54 \equiv (72.86, 75.94).$$



**Exemple:**

Soit un EAS  $\mathcal{Y} = \{(x_1, y_1), \dots, (x_n, y_n)\}$  de taille  $n = 132$  prélevé d'une population de taille  $N = 37,444$ .

Lors d'une étude préalable, il est déterminé que  $\mu_X \approx 70.3$ .

Donner un I.C. de  $\mu_Y$  à environ 95% par la différence si

$$\sum_{i=1}^n x_i = 9464.6, \quad \sum_{i=1}^n y_i = 14691.6,$$
$$\sum_{i=1}^n x_i^2 = 686773.2, \quad \sum_{i=1}^n x_i y_i = 1062186, \quad \sum_{i=1}^n y_i^2 = 1670194.$$

**Solution:** Nous avons déjà calculé

$$\bar{x} = 71.7, \bar{y} \approx 111.3, s_X^2 \approx 62.2, s_Y^2 \approx 267.3, s_{XY} \approx 67.2.$$

L'estimateur par la différence est alors

$$\hat{\mu}_{Y;D} = \bar{y} + (\mu_x - \bar{x}) = 111.3 + (70.3 - 71.7) = 109.9$$

et

$$\hat{B}_D \approx 2\sqrt{\frac{267.3 - 2(67.2) + 62.2}{132} \left(1 - \frac{132}{37444}\right)} = 2.427,$$

d'où

$$\text{IC}_D(\mu_Y; 0.95) \equiv 109.9 \pm 2.427.$$

(Comparer avec l'I.C. à la page 55.)

**Exemple:**

Donner un intervalle de confiance à 95% de l'espérance de vie moyenne par pays en 2011 en utilisant un EAS de taille  $n = 20$  et la méthode de la différence avec le logarithme du PNB (avec  $\mu_X = 7.84$ ).

**Solution:** On commence par préparer l'ensemble de données.

---

```
> gapminder.QRD <- gapminder %>% filter(year==2011) %>%  
  select(life_expectancy,gdp,population)  
> gapminder.QRD <- gapminder.QRD[complete.cases(gapminder.QRD),]  
> N=nrow(gapminder.QRD)  
> gapminder.QRD <- gapminder.QRD %>% mutate(lgdppc=log(gdp/population))  
> (mu.X = mean(gapminder.QRD[, "lgdppc"]))
```

---

```
[1] 7.842661
```

On prépare ensuite un échantillon de taille  $n = 20$ , et on calcule:

---

```
> n=20
> set.seed(1234567)
> index = sample(1:N,n, replace=FALSE)
> ech = gapminder.QRD[index,c("life_expectancy","lgdppc")]
> d = ech$life_expectancy - ech$lgdppc

# moyennes d'échantillon
> (y.bar = mean(ech$life_expectancy))
> (x.bar = mean(ech$lgdppc))
> (d.bar = mean(d))
> (s2.d = var(d))
```

---

```
[1] 72.145 8.288675 63.85633 62.24335
```

La pente de la relation entre  $Y$  et  $X$  ne semble pas être 1...

---

```
# marge d'erreur
> B = 2*sqrt(s2.d/n*(1-n/N))

# estime ponctuel
> hat.mu.Y.D = y.bar + (mu.X-x.bar)

# intervalle de confiance
> c(hat.mu.Y.D-B,hat.mu.Y.D+B)
```

---

```
[1] 68.38739 75.01059
```

Pourtant, l'I.C. à 95% pour l'espérance de vie moyenne par pays contient la valeur réelle,  $\mu_Y = 70.95$ !

À l'instar des autres méthodes, on peut déterminer la taille de l'échantillon requise afin d'atteindre une certaine marge d'erreur sur l'estimation.

Lorsque l'on cherche à estimer  $\mu_Y$  et  $\tau_Y$  par la différence, nous obtenons


$$B_{\mu;D} \approx 2\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\mu;D} = \frac{N\sigma_D^2}{(N-1)D_{\mu;D} + \sigma_D^2};$$
$$B_{\tau;D} \approx 2N\sqrt{\frac{\sigma_D^2}{n}\left(\frac{N-n}{N-1}\right)} \iff n_{\tau;D} = \frac{N\sigma_D^2}{(N-1)D_{\tau;D} + \sigma_D^2},$$

où

$$D_{\mu;D} = \frac{B_{\mu;D}^2}{4} \quad \text{et} \quad D_{\tau;D} = \frac{B_{\tau;D}^2}{4N^2}.$$



Puisque nous ne connaissons pas  $\sigma_D^2$  en général, on prélève souvent un échantillon préliminaire de petite taille et on utilise la **variance empirique**  $s_D^2$  en tant qu'estimateur de  $\sigma_D^2$ .

 Même si les manipulations formelles peuvent toujours s'effectuer, **l'estimé risque de ne pas être valide si la relation entre les variables  $X$  et  $Y$  n'est pas linéaire et de pente  $\approx 1$ .**

## 4.5 – Comparaisons

Nous avons déjà comparé les marges d'erreur sur l'estimation dans le cas de l'EAS, du STR (AP), et du STR (AN), et discuté de contextes dans lequel on peut s'attendre à ce qu'un STR soit préférable à un EAS, ou un STR (AN) préférable à un STR (AP).

Que peut-on dire au sujet de l'estimation par le quotient, par la régression, et par la différence?

En comparaison avec un EAS, et entre eux-mêmes?

### 4.5.1 – Comparaison entre EAS et la méthode du quotient

Dans quel contexte peut-on s'attendre à ce que la méthode d'estimation par le quotient donne de “bons” résultats?

De toute évidence, la relation entre  $Y$  et  $X$  doit au moins être **linéaire** et **passer par l'origine**, c'est-à-dire

$$y_i = \beta x_i + \varepsilon_i, \quad i = 1, \dots, n.$$

On suppose en général que les observations  $\{x_i > 0\}$  sont fixes, et que les termes d'erreur  $\{\varepsilon_i\}$  sont indépendant les uns des autres, avec

$$E(\varepsilon_i) = 0 \quad \text{et} \quad V(\varepsilon_i) = f(x_i)\sigma^2 > 0.$$

La question devient la suivante: quelle forme doit prendre  $f(x_i)$  afin que la solution des moindres carrés  $\hat{\beta}$  soit **exactement** l'estimateur  $r$  du quotient  $R$ ?

Si l'on pose

$$\underbrace{\frac{y_i}{\sqrt{f(x_i)}}}_{y'_i} = \beta \underbrace{\frac{x_i}{\sqrt{f(x_i)}}}_{x'_i} + \underbrace{\frac{\varepsilon_i}{\sqrt{f(x_i)}}}_{\varepsilon'_i}, \quad i = 1, \dots, n,$$

nous obtenons

$$E(\varepsilon'_i) = \frac{1}{\sqrt{f(x_i)}} E(\varepsilon) = 0 \quad \text{et} \quad V(\varepsilon'_i) = \frac{1}{f(x_i)} V(\varepsilon_i) = \frac{f(x_i)\sigma^2}{f(x_i)} = \sigma^2,$$

et les hypothèses du problème des moindres carrés sont alors satisfaites, et on trouve l'estimateur de  $\beta$  en minimisant

$$Q(\beta) = \sum_{i=1}^n (\varepsilon'_i)^2 = \sum_{i=1}^n (y'_i - \beta x'_i)^2 = \sum_{i=1}^n \frac{1}{f(x_i)} (y_i - \beta x_i)^2;$$

puisque

$$Q'(\beta) = -2 \sum_{i=1}^n \frac{x_i}{f(x_i)} (y_i - \beta x_i),$$

cela revient à résoudre

$$0 = \sum_{i=1}^n \frac{x_i}{f(x_i)} (y_i - \hat{\beta} x_i) \iff 0 = \sum_{i=1}^n \left( \frac{x_i y_i}{f(x_i)} - \hat{\beta} \frac{x_i^2}{f(x_i)} \right) \iff \hat{\beta} = \frac{\sum_{i=1}^n \frac{x_i y_i}{f(x_i)}}{\sum_{i=1}^n \frac{x_i^2}{f(x_i)}}.$$

Si  $\frac{x_i}{f(x_i)} = k > 0$  pour tout  $i = 1, \dots, n$ , l'estimateur  $\hat{\beta}$  devient

$$\hat{\beta} = \frac{k \sum_{i=1}^n y_i}{k \sum_{i=1}^n x_i} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i} = r.$$

Ainsi, lorsque la variance de  $Y$  le long de la droite  $Y = \beta X$  est

$$V(y_i) = V(\beta x_i + \varepsilon_i) = V(\varepsilon_i) = x_i \sigma^2$$

(i.e. **la variance de  $Y$  est proportionnelle à  $X$** ), l'estimateur  $r$  du quotient  $R$  est exactement la solution des moindres carrés,  $\hat{\beta} = r$ , et on peut s'attendre à ce que l'estimation par le quotient donne de bons résultats.

Bien sûr, on peut se servir de la méthode de l'estimation par le quotient avec un EAS  $\mathcal{Y}$  afin d'obtenir un estimé  $\hat{\mu}_{Y;R}$  de  $\mu_Y$ , même si  $V(\varepsilon) \neq x\sigma^2$ .

Nous avons déjà déterminé la variance de cet estimateur:

$$\begin{aligned} V(\hat{\mu}_{Y;R}) &= V(r\mu_X) = \mu_X^2 V(r) \approx \frac{1}{n}(\sigma_Y^2 + R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y) \left( \frac{N-n}{N-1} \right) \\ &= \underbrace{\frac{\sigma_Y^2}{n} \left( \frac{N-n}{N-1} \right)}_{V(\bar{y}_{\text{EAS}})} + \frac{R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y}{n} \left( \frac{N-n}{N-1} \right). \end{aligned}$$

Ainsi,  $V(\bar{y}_{\text{EAS}}) \gg V(\hat{\mu}_{Y;R})$  si et seulement si  $R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \ll 0$ , c'est-à-dire si

$$\rho \gg \frac{R\sigma_X}{2\sigma_Y} = \frac{\mu_Y\sigma_X}{2\mu_X\sigma_Y} = \frac{1}{2} \cdot \frac{\text{CV}_X}{\text{CV}_Y}.$$

## 4.5.2 – Comparaison entre EAS et la méthode de la régression

Nous avons déjà déterminé la variance de l'estimateur  $\hat{\mu}_{Y;L}$  de  $\mu_Y$ :

$$\begin{aligned} V(\hat{\mu}_{Y;L}) &\approx (1 - \rho^2) \frac{\sigma_Y^2}{n} \left( \frac{N - n}{N - 1} \right) = \underbrace{\frac{\sigma_Y^2}{n} \left( \frac{N - n}{N - 1} \right)}_{V(\bar{y}_{EAS})} - \rho^2 \cdot \frac{\sigma_Y^2}{n} \left( \frac{N - n}{N - 1} \right) \\ &= (1 - \rho^2) V(\bar{y}_{EAS}). \end{aligned}$$

Ainsi,  $V(\hat{\mu}_{Y;L}) \ll V(\bar{y}_{EAS})$  lorsque  $(1 - \rho^2) V(\bar{y}_{EAS}) \ll V(\bar{y}_{EAS})$ , c'est à dire

$$1 - \rho^2 \ll 1 \iff 0 \ll |\rho| \leq 1.$$



### 4.5.3 – Comparaison entre EAS et la méthode de la différence

Nous avons déjà déterminé la variance de l'estimateur  $\hat{\mu}_{Y;D}$  de  $\mu_Y$ :

$$\begin{aligned} V(\hat{\mu}_{Y;D}) &= \frac{\sigma_Y^2 - 2\rho\sigma_X\sigma_Y + \sigma_X^2}{n} \left( \frac{N-n}{N-1} \right) \\ &= \underbrace{\frac{\sigma_Y^2}{n} \left( \frac{N-n}{N-1} \right)}_{V(\bar{y}_{\text{EAS}})} + \frac{\sigma_X^2 - 2\rho\sigma_X\sigma_Y}{n} \left( \frac{N-n}{N-1} \right). \end{aligned}$$

Ainsi,  $V(\hat{\mu}_{Y;D}) \ll V(\bar{y}_{\text{EAS}})$  lorsque  $\sigma_X^2 - 2\rho\sigma_X\sigma_Y \ll 0 \iff \sigma_X^2 \ll 2\sigma_{XY}$ .

### 4.5.4 – Comparaison entre la méthode du quotient, la méthode de la régression, et la méthode de la différence

Pour chacun des estimateurs  $\hat{\mu}_{Y;\alpha}$ ,  $\alpha \in \{R, L, D\}$ , nous avons montré que la variance d'échantillonnage prend la forme (approximative)

$$V(\hat{\mu}_{Y;\alpha}) \approx V(\bar{y}_{\text{EAS}}) + \frac{A_\alpha}{n} \left( \frac{N-n}{N-1} \right),$$

où

$$A_\alpha = \begin{cases} R^2 \sigma_X^2 - 2R\rho\sigma_X\sigma_Y, & \alpha = R \\ -\rho^2 \sigma_Y^2, & \alpha = L \\ \sigma_X^2 - 2\rho\sigma_X\sigma_Y, & \alpha = D \end{cases}$$

En général,  $V(\hat{\mu}_Y; \alpha) \ll V(\hat{\mu}_Y; \gamma)$  si et seulement si  $A_\alpha \ll A_\gamma$ ; ce sont ces termes qu'il faut comparer.

Nous obtenons, par exemple,

$$\begin{aligned}
 V(\hat{\mu}_Y; R) \gg V(\hat{\mu}_Y; L) &\iff R^2 \sigma_X^2 - 2R\rho\sigma_X\sigma_Y \gg -\rho^2 \sigma_Y^2 \\
 &\iff R^2 \sigma_X^2 - 2R\rho\sigma_X\sigma_Y + \rho^2 \sigma_Y^2 \gg 0 \\
 &\iff (R\sigma_X - \rho\sigma_Y)^2 \gg 0 \iff |R\sigma_X - \rho\sigma_Y| \gg 0 \\
 &\iff R \gg \rho \frac{\sigma_Y}{\sigma_X} = \hat{\beta} \quad \text{ou} \quad R \ll \hat{\beta}
 \end{aligned}$$

Toutes choses étant égales, l'estimateur par la régression est préférable à l'estimateur par le quotient (au sens de la marge d'erreur) lorsque **le quotient est très différent de la pente de la droite de régression**.

Dans le même ordre d'idée,

$$\begin{aligned} V(\hat{\mu}_{Y;D}) \gg V(\hat{\mu}_{Y;L}) &\iff \sigma_X^2 - 2\rho\sigma_X\sigma_Y \gg -\rho^2\sigma_Y^2 \\ &\iff \sigma_X^2 - 2\rho\sigma_X\sigma_Y + \rho^2\sigma_Y^2 \gg 0 \\ &\iff (\sigma_X - \rho\sigma_Y)^2 \gg 0 \iff |\sigma_X - \rho\sigma_Y| \gg 0 \\ &\iff 1 \gg \rho \frac{\sigma_Y}{\sigma_X} = \hat{\beta} \quad \text{ou} \quad 1 \ll \hat{\beta} \end{aligned}$$

Toutes choses étant égales, l'estimateur par la régression est préférable à l'estimateur par la différence (au sens de la marge d'erreur) lorsque **la pente de la droite de régression prend une valeur éloignée de 1**.

L'estimateur par la régression est toujours **au moins aussi performant que les deux autres** puisque ces derniers en sont des cas spéciaux.

Finalement, on peut aussi comparer les estimateurs par le quotient et par la différence:

$$\begin{aligned} V(\hat{\mu}_{Y;R}) \gg V(\hat{\mu}_{Y;D}) &\iff R^2\sigma_X^2 - 2R\rho\sigma_X\sigma_Y \gg \sigma_X^2 - 2\rho\sigma_X\sigma_Y \\ &\iff |R| \neq 1 \quad \text{et} \quad \sigma_X^2 \gg \frac{2}{R+1}\sigma_{XY} \end{aligned}$$

et

$$V(\hat{\mu}_{Y;D}) \gg V(\hat{\mu}_{Y;R}) \iff |R| \neq 1 \quad \text{et} \quad \sigma_X^2 \ll \frac{2}{R+1}\sigma_{XY}$$

Autrement, les variances sont de magnitude relativement semblables.