

# *Cours de Statistiques Inférentielles*

CQLS : [cqls@upmf-grenoble.fr](mailto:cqls@upmf-grenoble.fr)

9 février 2018

# Plan

## 1 Estimation : Obtention et Qualité

# Problématique du Salaire Juste

## Enoncé

- Une équipe de sociologues propose de réunir un comité d'experts pour la création d'un indicateur, appelé **Salaire Juste**, mesuré pour toute personne active et qui permettra de transformer les ressources individuelles réelles (souvent mesurées par un salaire) en tenant compte de critères aussi importants que les ressources locales, le partage de ces ressources, la pénibilité du travail, le niveau d'expérience, d'expertise et bien d'autres encore.
- Cet indicateur est conçu de sorte qu'en théorie il devrait être équivalent (en fait égale à une valeur étalon 100) pour tout personne active dans le monde.
- Après quelques mois de travail, un premier prototype (très perfectible) du **Salaire Juste** est élaboré par la fine équipe d'experts.

# Problématique du Salaire Juste

## *Critère de pays civilisé*

Les sociologues s'accordent à dire qu'un pays peut se dire non civilisé si :

- ❶ **Discrimination Mondiale** : le Salaire Juste moyen dans le pays est très supérieur à la valeur 100 de base. Un Salaire Juste moyen excédant un seuil de 150 est considéré comme intolérable.
- ❷ **Discrimination Intérieure** : les Salaires Justes dans le pays sont très dispersés. La variance des Salaires Justes dans le pays supérieur à 30 est considérée comme excessive et donc anormale.

# Problématique du Salaire Juste

## Mesures de discrimination

Les experts sont aussi conseillés par des statisticiens pour proposer les mesures de discrimination au niveau du pays et mondialement.  $\mathcal{Y}_i^J$  désigne le Salaire Juste du  $i^{\text{ème}}$  individu parmi les  $N$  personnes actives du pays.  $Y^J$  correspond au Salaire Juste d'un individu choisi au hasard.

❶ **Discrimination Mondiale** : le Salaire Juste moyen s'écrit :

$$\mu^J = (\overline{\mathcal{Y}^J})_N = \frac{1}{N} \sum_{i=1}^N \mathcal{Y}_i^J = \mathbb{E}(Y^J)$$

❷ **Discrimination Intérieure** : la variance des Salaires Justes s'écrit :

$$\sigma_J^2 = \left( \overleftrightarrow{(\mathcal{Y}^J)}_N \right)^2 = \frac{1}{N} \sum_{i=1}^N \left( \mathcal{Y}_i^J - (\overline{\mathcal{Y}^J})_N \right)^2 = \mathbb{V}ar(Y^J)$$

## Estimation

Les paramètres (d'intérêt)  $\mu_J$  et  $\sigma_J^2$  (appelé,  $\theta^\bullet$  dans un cadre général) sont donc supposés **inconnus** car la taille  $N$  de la population est trop grande. Proposez les estimations ?

## Estimation

Les paramètres (d'intérêt)  $\mu^J$  et  $\sigma_J^2$  (appelé,  $\theta^\bullet$  dans un cadre général) sont donc supposés **inconnus** car la taille  $N$  de la population est trop grande. Proposez les estimations ?

Future estimation  $\hat{\theta}^\bullet(\mathbf{Y})$  : (à partir d'un futur échantillon  $\mathbf{Y}$ )

$$\widehat{\mu}^J(\mathbf{Y}) := \overline{(Y)}_n := \frac{1}{n} \sum_{i=1}^n Y_i$$

et

$$\widehat{\sigma}_J^2(\mathbf{Y}) := \frac{1}{\textcolor{red}{n} - \textcolor{red}{1}} \sum_{i=1}^n \left( Y_i - \overline{(Y)}_n \right)^2$$

## Estimation

Les paramètres (d'intérêt)  $\mu^J$  et  $\sigma_J^2$  (appelé,  $\theta^\bullet$  dans un cadre général) sont donc supposés **inconnus** car la taille  $N$  de la population est trop grande. Proposez les estimations ?

Estimation  $\hat{\theta}^\bullet(\mathbf{y})$  du Jour J : (à partir de l'échantillon  $\mathbf{y}$  du Jour J)

$$\hat{\mu}^J(\mathbf{y}) := \overline{(y)}_n := \frac{1}{n} \sum_{i=1}^n y_i$$

et

$$\hat{\sigma}_J^2(\mathbf{y}) := \frac{1}{\textcolor{red}{n} - \textcolor{red}{1}} \sum_{i=1}^n \left( y_i - \overline{(y)}_n \right)^2$$



## Estimation

Les paramètres (d'intérêt)  $\mu^J$  et  $\sigma_J^2$  (appelé,  $\theta^\bullet$  dans un cadre général) sont donc supposés **inconnus** car la taille  $N$  de la population est trop grande. Proposez les estimations ?

Estimations potentielles  $\hat{\theta}^\bullet(\mathbf{y}_{[k]})$  : (à partir d'échantillons  $\mathbf{y}_{[k]}$ )

$$\widehat{\mu}^J(\mathbf{y}_{[k]}) := \overline{(y_{[k]})}_n := \frac{1}{n} \sum_{i=1}^n y_{i,[k]}$$

et

$$\widehat{\sigma}_J^2(\mathbf{y}_{[k]}) := \frac{1}{\textcolor{red}{n} - 1} \sum_{i=1}^n \left( y_{i,[k]} - \overline{(y_{[k]})}_n \right)^2$$

## Qualité

Quelles sont les qualités souhaitables pour l'estimation d'un paramètre d'intérêt ? Pouvez-vous les traduire à partir des  $m$  estimations potentielles ? (*N.B. : l'estimation du jour  $J$  est choisi parmi celles-ci*)

# Qualité

Quelles sont les qualités souhaitables pour l'estimation d'un paramètre d'intérêt ? Pouvez-vous les traduire à partir des  $m$  estimations potentielles ? (*N.B. : l'estimation du jour  $J$  est choisi parmi celles-ci*)

- **Autour du paramètre** : leur moyenne égale au paramètre inconnu  $\theta^\bullet$  (**A.M.P.** : estimateur sans biais)

$$\overline{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_m \simeq \overline{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_\infty = \mathbb{E}(\hat{\theta}^\bullet(\mathbf{Y})) = \theta^\bullet$$

- **Faible dispersion** : leur écart-type (ou variance) d'autant plus petit que  $n$  grandit (**A.M.P.** : estimateur convergent)

$$\overleftarrow{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_m \simeq \overleftarrow{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_\infty = \sigma(\hat{\theta}^\bullet(\mathbf{Y})) \xrightarrow{n \rightarrow +\infty} 0$$

Quelles sont les qualités souhaitables pour l'estimation d'un paramètre d'intérêt ? Pouvez-vous les traduire à partir des  $m$  estimations potentielles ? (*N.B. : l'estimation du jour  $J$  est choisi parmi celles-ci*)

- **Autour du paramètre** : leur moyenne égale au paramètre inconnu  $\theta^\bullet$  (**A.M.P.** : estimateur sans biais)

$$\overline{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_m \simeq \overline{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_\infty = \mathbb{E}(\hat{\theta}^\bullet(\mathbf{Y})) = \theta^\bullet$$

- **Faible dispersion** : leur écart-type (ou variance) d'autant plus petit que  $n$  grandit (**A.M.P.** : estimateur convergent)

$$\overleftarrow{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_m \simeq \overleftarrow{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_\infty = \sigma(\hat{\theta}^\bullet(\mathbf{Y})) \xrightarrow{n \rightarrow +\infty} 0$$

# Qualité

Quelles sont les qualités souhaitables pour l'estimation d'un paramètre d'intérêt ? Pouvez-vous les traduire à partir des  $m$  estimations potentielles ? (*N.B. : l'estimation du jour  $J$  est choisi parmi celles-ci*)

- **Autour du paramètre** : leur moyenne égale au paramètre inconnu  $\theta^\bullet$  (**A.M.P.** : estimateur sans biais)

$$\overline{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_m \simeq \overline{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_\infty = \mathbb{E}(\hat{\theta}^\bullet(\mathbf{Y})) = \theta^\bullet$$

- **Faible dispersion** : leur écart-type (ou variance) d'autant plus petit que  $n$  grandit (**A.M.P.** : estimateur convergent)

$$\overleftarrow{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_m \simeq \overleftarrow{(\hat{\theta}^\bullet(\mathbf{y}_{[.]})}_\infty = \sigma(\hat{\theta}^\bullet(\mathbf{Y})) \xrightarrow{n \rightarrow +\infty} 0$$

$\Rightarrow$  **Pb** : qualité d'estimation  $\sigma_{\hat{\theta}^\bullet} := \sigma(\hat{\theta}^\bullet(\mathbf{Y}))$  est un paramètre **inconnu** ! Peut-on espérer l'estimer à partir de l'échantillon  $\mathbf{y}$  ?

## Erreur standard

Les statisticiens (mathématiciens) proposent généralement l'estimation  $\hat{\theta}(\mathbf{y})$  d'un paramètre inconnu  $\theta^\bullet$  accompagnée de l'estimation  $\hat{\sigma}_{\theta^\bullet}(\mathbf{y})$  de sa qualité  $\sigma_{\theta^\bullet}$ .

(Voir le tableau dans votre caisse à outils pour la liste de toutes les erreurs standard associées aux différents paramètres!)

Pour illustrer comment cela est possible, étudions le paramètre moyenne  $\mu^\bullet$  :

$$\hat{\sigma}_{\mu^\bullet} = \frac{\sigma_{\bullet}}{\sqrt{n}}$$

## Erreur standard

Les statisticiens (mathématiciens) proposent généralement l'estimation  $\widehat{\theta}(\mathbf{y})$  d'un paramètre inconnu  $\theta^\bullet$  accompagnée de l'estimation  $\widehat{\sigma_{\theta^\bullet}}(\mathbf{y})$  de sa qualité  $\sigma_{\theta^\bullet}$ .

(Voir le tableau dans votre caisse à outils pour la liste de toutes les erreurs standard associées aux différents paramètres!)

Pour illustrer comment cela est possible, étudions le paramètre moyenne  $\mu^\bullet$  :

$$\widehat{\sigma_{\mu^\bullet}} = \frac{\sigma_{\bullet}}{\sqrt{n}} \text{ estimé par } \widehat{\sigma_{\mu^\bullet}}(\mathbf{Y}) = \frac{\widehat{\sigma_{\bullet}}(\mathbf{Y})}{\sqrt{n}}$$

**Objectif :** Etude de la loi de l'écart  $\hat{\theta}^\bullet(\mathbf{Y}) - \theta^\bullet$  (potentiellement, fort utile pour construction d'outils statistiques)

$\mathbf{Y}$	$\hat{\theta}^\bullet(\mathbf{Y})$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{Y})$
$\mathbf{y}[1]$	$\hat{\theta}^\bullet(\mathbf{y}[1])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[1])$
$\mathbf{y}[2]$	$\hat{\theta}^\bullet(\mathbf{y}[2])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[2])$
$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}[m-1]$	$\hat{\theta}^\bullet(\mathbf{y}[m-1])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[m-1])$
$\mathbf{y}[m]$	$\hat{\theta}^\bullet(\mathbf{y}[m])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[m])$
$\vdots$	$\vdots$	$\vdots$



**Problème :** Loi de  $\hat{\theta}^\bullet(\mathbf{Y}) - \theta^\bullet$  généralement inconnue car dépendant d'un paramètre de nuisance inconnu  $\sigma_{\hat{\theta}^\bullet}$ .

$\mathbf{Y}$	$\hat{\theta}^\bullet(\mathbf{Y})$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{Y})$	$\hat{\theta}^\bullet(\mathbf{Y}) - \theta^\bullet$
$\mathbf{y}[1]$	$\hat{\theta}^\bullet(\mathbf{y}[1])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[1])$	$\hat{\theta}^\bullet(\mathbf{y}[1]) - \theta^\bullet$
$\mathbf{y}[2]$	$\hat{\theta}^\bullet(\mathbf{y}[2])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[2])$	$\hat{\theta}^\bullet(\mathbf{y}[2]) - \theta^\bullet$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}[m-1]$	$\hat{\theta}^\bullet(\mathbf{y}[m-1])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[m-1])$	$\hat{\theta}^\bullet(\mathbf{y}[m-1]) - \theta^\bullet$
$\mathbf{y}[m]$	$\hat{\theta}^\bullet(\mathbf{y}[m])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[m])$	$\hat{\theta}^\bullet(\mathbf{y}[m]) - \theta^\bullet$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Loi ( $n \geq 30$ )	$\mathcal{N}(\theta^\bullet, \sigma_{\hat{\theta}^\bullet})$		$\mathcal{N}(0, \sigma_{\hat{\theta}^\bullet})$

# Mesure d'écart standardisé

**La solution :** La loi de l'écart standardisé (centrage et réduction)

$$\delta_{\hat{\theta}^\bullet, \theta^\bullet}(\mathbf{Y}) := \frac{\hat{\theta}^\bullet(\mathbf{Y}) - \theta^\bullet}{\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{Y})} \underset{\text{approx.}}{\rightsquigarrow} \mathcal{N}(0, 1)$$

$\mathbf{Y}$	$\hat{\theta}^\bullet(\mathbf{Y})$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{Y})$	$\hat{\theta}^\bullet(\mathbf{Y}) - \theta^\bullet$	$\delta_{\hat{\theta}^\bullet, \theta^\bullet}(\mathbf{Y})$
$\mathbf{y}[1]$	$\hat{\theta}^\bullet(\mathbf{y}[1])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[1])$	$\hat{\theta}^\bullet(\mathbf{y}[1]) - \theta^\bullet$	$\delta_{\hat{\theta}^\bullet, \theta^\bullet}(\mathbf{y}[1])$
$\mathbf{y}[2]$	$\hat{\theta}^\bullet(\mathbf{y}[2])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[2])$	$\hat{\theta}^\bullet(\mathbf{y}[2]) - \theta^\bullet$	$\delta_{\hat{\theta}^\bullet, \theta^\bullet}(\mathbf{y}[2])$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}[m-1]$	$\hat{\theta}^\bullet(\mathbf{y}[m-1])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[m-1])$	$\hat{\theta}^\bullet(\mathbf{y}[m-1]) - \theta^\bullet$	$\delta_{\hat{\theta}^\bullet, \theta^\bullet}(\mathbf{y}[m-1])$
$\mathbf{y}[m]$	$\hat{\theta}^\bullet(\mathbf{y}[m])$	$\widehat{\sigma_{\hat{\theta}^\bullet}}(\mathbf{y}[m])$	$\hat{\theta}^\bullet(\mathbf{y}[m]) - \theta^\bullet$	$\delta_{\hat{\theta}^\bullet, \theta^\bullet}(\mathbf{y}[m])$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Loi ( $n \geq 30$ )	$\mathcal{N}(\theta^\bullet, \sigma_{\hat{\theta}^\bullet})$		$\mathcal{N}(0, \sigma_{\hat{\theta}^\bullet})$	$\mathcal{N}(0, 1)$

## Mesure d'écart standardisé

**La solution :** La loi de l'écart standardisé (centrage et réduction)

$$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{Y}) := \frac{\widehat{\mu}^J(\mathbf{Y}) - \mu^J}{\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{Y})} \underset{\text{approx.}}{\rightsquigarrow} \mathcal{N}(0, 1) \text{ avec } \widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{Y}) := \frac{\widehat{\sigma}^J(\mathbf{Y})}{\sqrt{n}}$$

$\mathbf{Y}$	$\widehat{\mu}^J(\mathbf{Y})$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{Y})$	$\widehat{\mu}^J(\mathbf{Y}) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{Y})$
$\mathbf{y}[1]$	$\widehat{\mu}^J(\mathbf{y}[1])$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{y}[1])$	$\widehat{\mu}^J(\mathbf{y}[1]) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{y}[1])$
$\mathbf{y}[2]$	$\widehat{\mu}^J(\mathbf{y}[2])$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{y}[2])$	$\widehat{\mu}^J(\mathbf{y}[2]) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{y}[2])$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}[m-1]$	$\widehat{\mu}^J(\mathbf{y}[m-1])$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{y}[m-1])$	$\widehat{\mu}^J(\mathbf{y}[m-1]) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{y}[m-1])$
$\mathbf{y}[m]$	$\widehat{\mu}^J(\mathbf{y}[m])$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{y}[m])$	$\widehat{\mu}^J(\mathbf{y}[m]) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{y}[m])$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Loi ( $n \geq 30$ )	$\mathcal{N}(\mu^J, \sigma_{\widehat{\mu}^J})$		$\mathcal{N}(0, \sigma_{\widehat{\mu}^J})$	$\mathcal{N}(0, 1)$

## Mesure d'écart standardisé

**Application :** Salaire Juste avec population fictive fixée expérimentalement à  $\mu^J = 100$  et  $\sigma_J = 10$  avec taille d'échantillon  $n = 1000$ .

$\mathbf{Y}$	$\widehat{\mu}^J(\mathbf{Y})$	$\widehat{\sigma}_J(\mathbf{Y})$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{Y})$	$\widehat{\mu}^J(\mathbf{Y}) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{Y})$
y[1]	99.91	10.0231	0.317	-0.09	-0.29
y[2]	99.65	9.2615	0.2929	-0.35	-1.19
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y en R	mean(y)	sd(y)	seMean(y)	mean(y) - 100	(mean(y) - 100)/seMean(y)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y[9999]	100.2	9.9372	0.3142	0.2	0.63
y[10000]	99.94	9.7991	0.3099	-0.06	-0.21
mean	100.0043	10.034	0.3173	0.0043	-0.0294
sd	0.3178	0.63	0.0199	0.3178	1.0058

## Mesure d'écart standardisé

**Remarque :** Chaque ligne du tableau serait un résultat possible pour le **jour J**. Observez notamment la 1<sup>ère</sup> et 3<sup>ème</sup> colonnes (estimations des paramètres  $\mu^J$  et  $\sigma_J$  puis l'erreur standard pour  $\mu^J$ ) !

Y	$\widehat{\mu}^J(\mathbf{Y})$	$\widehat{\sigma}_J(\mathbf{Y})$	$\widehat{\sigma}_{\widehat{\mu}^J}(\mathbf{Y})$	$\widehat{\mu}^J(\mathbf{Y}) - \mu^J$	$\delta_{\widehat{\mu}^J, \mu^J}(\mathbf{Y})$
y[1]	99.91	10.0231	0.317	-0.09	-0.29
y[2]	99.65	9.2615	0.2929	-0.35	-1.19
⋮	⋮	⋮	⋮	⋮	⋮
y en R	mean(y)	sd(y)	seMean(y)	mean(y) - 100	(mean(y) - 100)/seMean(y)
⋮	⋮	⋮	⋮	⋮	⋮
y[9999]	100.2	9.9372	0.3142	0.2	0.63
y[10000]	99.94	9.7991	0.3099	-0.06	-0.21
mean	100.0043	10.034	0.3173	0.0043	-0.0294
sd	0.3178	0.63	0.0199	0.3178	1.0058

**Expérimentation** : Relation entre A.E.P. et A.M.P sur

$$\Delta := \delta_{\widehat{\mu}^J, \mu^J}(\mathbf{Y}) := \frac{\widehat{\mu}^J(\mathbf{Y}) - \mu^J}{\widehat{\sigma}_{\mu^J}(\mathbf{Y})} \overset{\text{approx.}}{\rightsquigarrow} \mathcal{N}(0, 1)$$

$\overline{(\delta_{[\cdot]} < -3)}_m$	$\overline{(\delta_{[\cdot]} \in [-3, -1.5])}_m$	$\overline{(\delta_{[\cdot]} \in [-1.5, -0.5])}_m$	$\overline{(\delta_{[\cdot]} \in [-0.5, 0.5])}_m$
0.23%	7.4%	23.94%	37.55%
$\mathbb{P}(\Delta < -3)$	$\mathbb{P}(\Delta \in [-3, -1.5])$	$\mathbb{P}(\Delta \in [-1.5, -0.5])$	$\mathbb{P}(\Delta \in [-0.5, 0.5])$
0.13%	6.55%	24.17%	38.29%

$\overline{(\delta_{[\cdot]} \in [0.5, 1.5])}_m$	$\overline{(\delta_{[\cdot]} \in [1.5, 3])}_m$	$\overline{(\delta_{[\cdot]} \geq 3)}_m$	$\overline{(\delta_{[\cdot]})}_m$	$\overleftrightarrow{(\delta_{[\cdot]})}_m$
25.05%	5.73%	0.1%	-0.0294	1.0058
$\mathbb{P}(\Delta \in [0.5, 1.5])$	$\mathbb{P}(\Delta \in [1.5, 3])$	$\mathbb{P}(\Delta \geq 3)$	$\mathbb{E}(\Delta)$	$\sigma(\Delta)$
24.17%	6.55%	0.13%	0	1