MAT4681 - Statistique pour les sciences

Arthur Charpentier

04 - Moyenne, variance (et rappels de maths) # 3

été 2022

Average, mean, median, mode, etc

Moyenne (empirique) / empirical mean / average

Pour un échantillon $\{y_1, \cdots, y_n\}$, la moyenne est $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$

Pour information, en python

```
> import statistics
> y = [1, 2, 3, 4, 5, 6]
> print(statistics.mean(y))
3.5
```

et en R,

```
> y = c(1, 2, 3, 4, 5, 6)
> mean(y)
[1] 3.5
```

Average, mean, median, mode, etc ★★★

Moyenne (empirique) / empirical mean / average

$\overline{y}$ est la solution de $\overline{y} = \underset{m\in\mathbb{R}}{\text{argmin}}\left\{\sum_{i=1}^n (y_i - m)^2\right\}$

Preuve: soit $\boldsymbol{y} = \{y_1, \cdots, y_n\}$, posons

$$g(m) = \sum_{i=1}^n (y_i - m)^2$$

$$\frac{\partial g(m)}{\partial m} = \frac{\partial}{\partial m}\sum_{i=1}^n (y_i - m)^2 = \sum_{i=1}^n \frac{\partial}{\partial m}(y_i - m)^2 = \sum_{i=1}^n -2(y_i - m)$$

Condition du premier ordre $\left.\frac{\partial g(m)}{\partial m}\right|_{m=m^\star} = 0$,

$$\sum_{i=1}^n (y_i - m^\star) = 0 \text{ si et seulement si } \sum_{i=1}^n y_i = n \cdot m^\star \text{ i.e. } m^\star = \overline{y}$$

Average, mean, median, mode, etc

Espérance mathématique

La moyenne est la version empirique de l'espérance d'une variable aléatoire,

$$\mathbb{E}(X) = \sum_x x\mathbb{P}[X = x] \text{ si } \sum_x |x|\mathbb{P}[X = x] < \infty$$

$$\mathbb{E}(X) = \int xf(x)dx \text{ si } \int |x|f(x)dx < \infty$$

Exemple: a coin has *heads* with probability p. Let $x = \mathbf{1}(heads)$,

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1 - p) = p$$

Exemple: if X is uniform over $[0, 1]$, $f(x) = \mathbf{1}_{[0,1]}(x)$

$$\mathbb{E}(X) = \int_0^1 x1dx = \frac{1}{2}$$

Average, mean, median, mode, etc

Linéarité de l'espérance mathématique

For all X and Y such that $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}[Y], \ \forall a, b \in \mathbb{R}.$$

$$\mathbb{E}(X_1 + \cdots + X_k) = \mathbb{E}(X_1) + \cdots \mathbb{E}(X_k), \ \forall X_1, \cdots, X_k$$

Example: toss n coins, of bias p, X is the number of heads

$$\mathbb{E}(X) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \cdots \mathbb{E}(X_n) = np$$

Average, mean, median, mode, etc

$$\mathbb{E}(X) = \sum_x x\mathbb{P}[X = x] = \sum_x xf(x) \text{ ou } \int xf(x)dx$$

$$\mathbb{E}(\psi(X)) = \sum_x \psi(x)\mathbb{P}[X = x] = \sum_x \psi(x)f(x) \text{ ou } \int \psi(x)f(x)dx$$

Exemple: pour une loi $\mathcal{N}(0,1)$, que vaut $\mathbb{E}[\cos[X]]$?

$$\mathbb{E}[\cos[X]] = \int_{-\infty}^{+\infty} \cos(x)\varphi(x)dx$$

```
> f = function(x) cos(x)*dnorm(x,0,1)
> integrate(f,-Inf,Inf)
0.6065307 with absolute error < 7.2e-08
> log(integrate(f,-Inf,Inf)$value)
[1] -0.5
```

Moyennes I

La moyenne est très sensible aux valeurs aberrantes (ou extrêmes, donc très grandes ou très petites)

```
> x = c(11, 10, 8, 10, 12, 14, 16, 7, 6, 9, 20, 10, 8,
     12, 13, 9, 10, 10, 10, 5, 10, 12, 9, 9, 10, 13)
> mean(x)
[1] 10.5
> x = c(11, 10, 8, 10, 12, 14, 16, 7, 6, 9, 20, 10, 8,
     12, 13, 9, 10, 10, 10, 100, 10, 12, 9, 9, 10, 13)
> mean(x)
[1] 14.15385
```

Un autre mesure robuste est la moyenne tronquée: elle consiste à calculer la moyenne arithmétique, mais en en enlevant une certaine proportion des observations en haut et en bas de la distribution.

Par exemple, la moyenne tronquée à 10% consiste à enlever les 10% des observations les plus grandes, et 10observations les plus petites.

Moyennes II

Soit $x_1, x_2, \cdots, x_n$ un échantillon ordonné, $x_1 \leq x_2 \leq \cdots \leq x_n$

Moyenne tronquée (trimmed average)

La moyenne tronquée de niveau $\alpha \in [0,1]$ est

$$\frac{1}{n-2k}\sum_{k+1}^{n-k} x_i \text{ où } k = \lfloor \alpha n \rfloor$$

```
> mean(Davis$height)
[1] 170.565
> mean(Davis$height, trim=.1)
[1] 170.3625
```

avec ici $\alpha = 10\%$.

Moyennes III

Pour $n = 10$ observations, la moyenne (régulière) est

$$\frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}}{10}$$

La moyenne olympique est obtenue en tronquant à $\alpha = 1/n$

Moyenne olympique (Olympic average)

$$\frac{x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9}{8}$$

alors que la moyenne de Windsor remplace x_1 par x_2

Moyenne de Windsor (Winsorized average)

$$\frac{\overbrace{x_2 + x_2} + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + \overbrace{x_9 + x_9}}{10}$$

Nonlinear transformation & Jensen Inequality ★★★

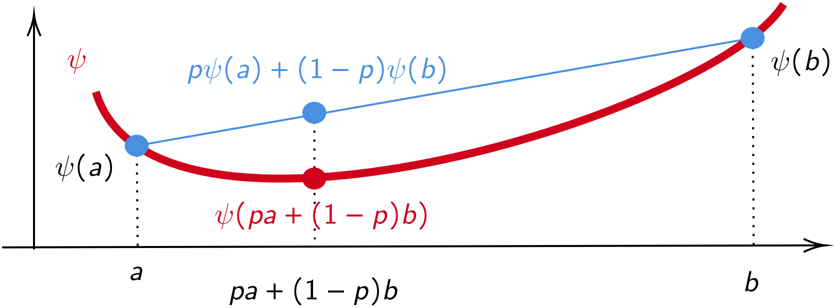
Let $\psi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(\psi(X)) = \sum_x \psi(x)\mathbb{P}[X = x] \text{ or } \int \psi(x)f(x)dx \neq \psi\big(\mathbb{E}(X)\big)$$

Example if X takes values in $\{a, b\}$, with probability p and $1 - p$,

$$\mathbb{E}(\psi(X)) = \psi(a)p + \psi(b)(1 - p)$$

If ψ is a convex function, $\mathbb{E}(\psi(X)) \geq \psi\big(\mathbb{E}(X)\big)$



Nonlinear transformation & Jensen Inequality ⭐⭐⭐

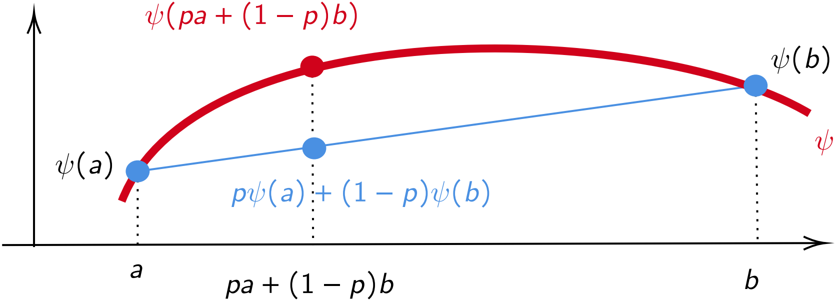
Let $\psi : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}(\psi(X)) = \sum_x \psi(x)\mathbb{P}[X = x] \text{ or } \int \psi(x)f(x)dx \neq \psi\big(\mathbb{E}(X)\big)$$

Example if X takes values in $\{a, b\}$, with probability p and $1 - p$,

$$\mathbb{E}(\psi(X)) = \psi(a)p + \psi(b)(1 - p)$$

If ψ is a concave function, $\mathbb{E}(\psi(X)) \leq \psi\big(\mathbb{E}(X)\big)$



St Petersburg's Paradox

As we will see (law of large numbers) if x_i are realizations of random variables X_i (with identical expected value μ), $\overline{x} \to \mu$ as $n \to \infty$.

A fair coin is tossed at each stage. The initial stake begins at 2 dollars and is doubled every time heads appears. The first time tails appears, the game ends and the player wins whatever is in the pot. Let X denote the gain.

$$\mathbb{E}(X) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \frac{1}{16} \cdot 16 + \cdots = 1 + 1 + 1 + 1 + \cdots = +\infty$$

the expected value is infinite (but the average always exists)

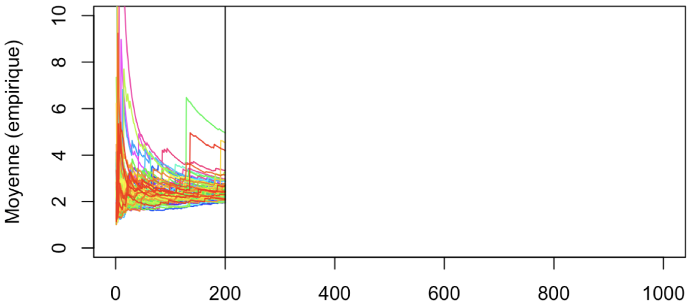
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha - 1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 1.7$



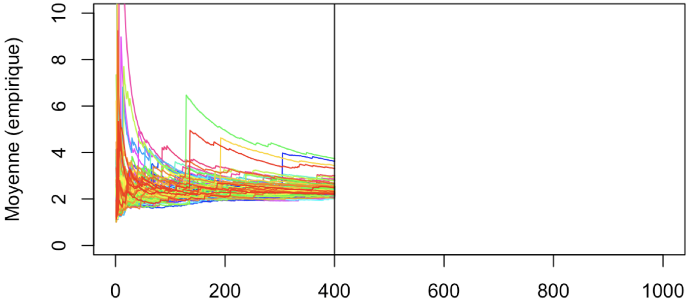
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha - 1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 1.7$



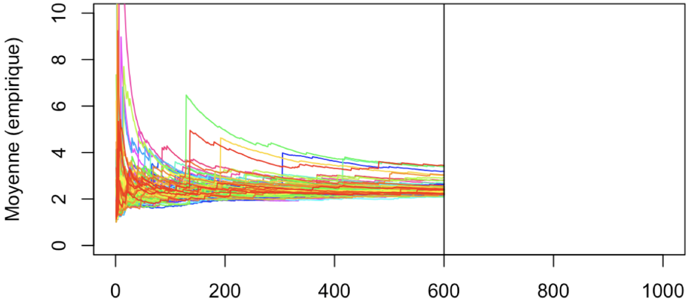
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha-1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 1.7$



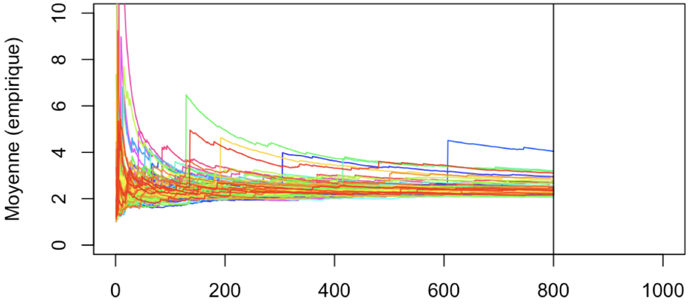
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha - 1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 1.7$



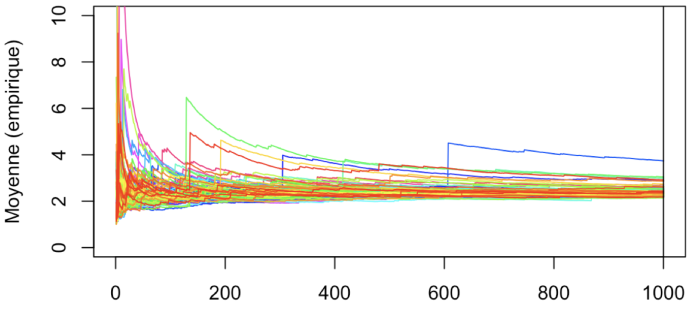
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha - 1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 1.7$



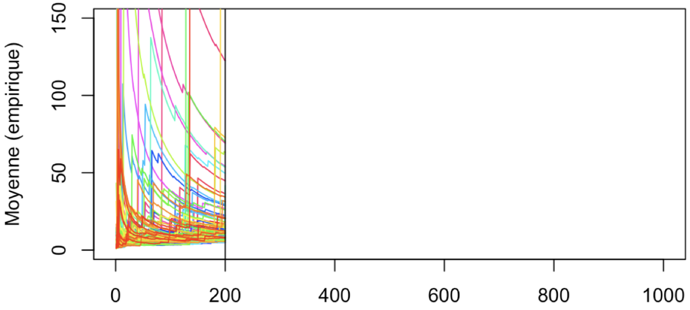
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha-1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 0.9$



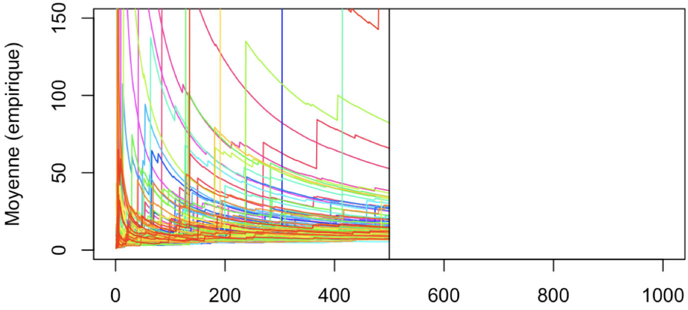
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha - 1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 0.9$



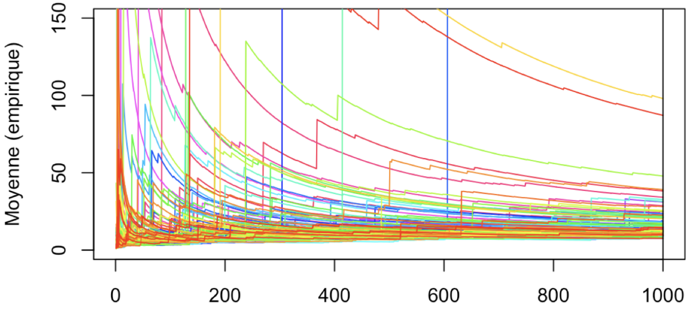
Average, mean, median, mode, etc

Example Loi de Pareto, $F(x) = 1 - x^{-\alpha}$ pour $x \geq 1$

La densité est $f(x) = \dfrac{\alpha}{x^{\alpha+1}}$ et l'espérance

$$\mathbb{E}[X] = \int_{x_m}^{\infty} \frac{x\alpha}{x^{\alpha+1}} dx = \begin{cases} \dfrac{\alpha}{\alpha - 1} & \text{si } \alpha > 1 \\ \infty & \text{si } \alpha \leq 1 \end{cases}$$

mais $\overline{x} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$ est toujours fini ! Exemple pour $\alpha = 0.9$



Average, mean, median, mode, quantiles etc

Quantile

Pour une fdr F,

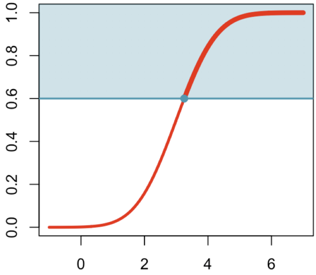
$$Q(p) = \inf\{x \in \mathbb{R} : p \leq F(x)\}$$

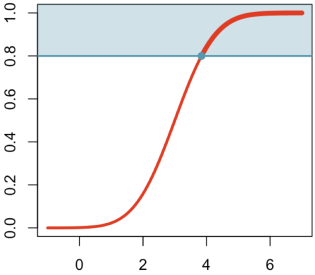
Le quantile est la seule fonction telle que

$$Q(p) \leq x \text{ si et seulement si } p \leq F(x)$$

Si F est continue et strictement croissante $Q(p) = F^{-1}(p)$.
Q est l'inverse à gauche:

$$Q(F(X)) = X$$





Average, mean, median, mode, quantiles etc

Intégrale de la fonction quantile

Si l'espérance d'une variable de loi F existe,

$$\int_0^1 Q(p)dp = \int_{-\infty}^{\infty} xf(x)dx = \mathbb{E}[X]$$

Preuve : par changement de variable $p = F(x)$,
$dp = F'(x)dx = f(x)dx$,

$$\int_0^1 F^{-1}(p)dp = \int_{-\infty}^{\infty} xf(x)dx = \mathbb{E}[X]$$

On peut aussi écrire

$$\mathbb{E}[X] = \mathbb{E}\big[(F^{-1}(U)\big] = \int_0^1 F^{-1}(u)du$$

où U suit une loi uniforme.

Average, mean, median, mode, quantiles etc

Exemple: pour une loi $\mathcal{N}(\mu, 1)$,

```
> mu = 7.3
> f = function(x) x*dnorm(x,mu,1)
> integrate(f,-Inf,Inf)
7.3 with absolute error < 3.3e-06
> g = function(p) qnorm(p,mu,1)
> integrate(g,0,1)
7.3 with absolute error < 8.1e-14
```

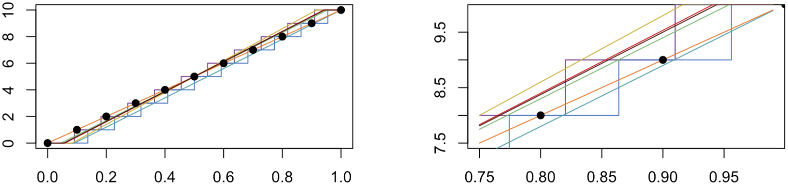
Définir les quantiles empiriques est plus compliqué

```
> ?quantile
> quantile(0:10,.95,type=7)
95%
9.5
> quantile(0:10,.95,type=3)
95%
  9
```

Average, mean, median, mode, quantiles etc

Considérons l'échantillon, $\boldsymbol{x} = \{1, 2, 3, \cdots, 9, 10\}$

```
> quantile (0:10 ,.9 , type =7)
90%
  9
```



Average, mean, median, mode, quantiles etc

Quantile empirique (1)

Notons $\{x_{(i)}\}$ une version ordonnée de $\{x_i\}$, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.Le quantile empirique de niveau $p \in (0,1)$ est

$$\widehat{q}_p = (1-f)x_{(k)} + fx_{(k+1)}$$

où $k = \lceil np \rceil$ et $f = n\alpha - \lfloor np \rfloor$.

Quantile empirique (2)

Étant donné $\{x_i\}$, si $\widehat{F}$ est la fonction de répartition empirique associée, on peut poser

$$\widetilde{q}_p = \widehat{F}^{-1}(p) = x_{(k)} \text{ où } k = \lceil np \rceil$$

Average, mean, median, mode, etc

The average is very sensitive to outliers and extremal values.

Médiane

Notons $\{x_{(i)}\}$ une version ordonnée de $\{x_i\}$, $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. On appelle médiane $Q(1/2)$, et sa version empirique est

$$md(x) = \begin{cases} x_{((n+1)/2)} & \text{si } n \text{ pair} \\ \dfrac{1}{2}\big(x_{(n/2)} + x_{(n/2+1)}\big) & \text{si } n \text{ impair} \end{cases}$$

(50% observations are smaller/larger)

Note that $md(x) \in \underset{m\in\mathbb{R}}{\text{argmin}}\left\{\sum_{i=1}^{n} |x_i - m|\right\}$

Average & Paradoxes

See also Will Rogers phenomenon,
"*Quand les Oklahoma ont quitté l'Oklahoma pour la Californie, ils ont augmenté le niveau d'intelligence moyen des deux États*"

$$\{1,2,3,4,5\}\ \{6,7,8,9,10\}$$

$$\{1,2,3,4,5,6\}\ \{7,8,9,10\}$$

See The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer for real implications

Variance

Given a sample $\boldsymbol{x} = \{x_1, \cdots, x_n\}$, $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$

```
> x = 1:6
> sum( (x-mean(x))^2 )/5
[1] 3.5
> var(x)
[1] 3.5
> sd(x)
[1] 1.870829
```

```
> import statistics
> x = [1, 2, 3, 4, 5, 6]
> print(statistics.variance(x))
3.5
> print(statistics.stdev(x))
1.8708286933869707
```

$s = \sqrt{s^2}$ is stdev($\boldsymbol{x}$) (standard deviation)

Dispersion, variance, standard deviation

Variance empirique (version 1)

Étant donné un échantillon $x_1, \cdots, x_n$, on appelle variance empirique $s^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$

Variance empirique (version 2)

Étant donné un échantillon $x_1, \cdots, x_n$, on appelle variance empirique $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$

Dispersion, variance, standard deviation

Variance empirique (version 2)

Étant donné un échantillon $x_1, \cdots, x_n$, on appelle variance empirique $s^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2$

Note C'est celle calculée par la plupart des logiciels

```
> x = 1:6
> var(x)
[1] 3.5
> sum((x-mean(x))^2)/5
[1] 3.5
> sum((x-mean(x))^2)/6
[1] 2.916667
```

C'est la version empirique de la variance ...

Dispersion, variance, standard deviation

Variance

On appelle variance d'une variable aléatoire

$$\text{Var}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

(à condition que $\mathbb{E}[X^2] < \infty$).

Covariance

On appelle covariance d'un couple de variable aléatoire

$$\text{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\big] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

(à condition que $\mathbb{E}[X^2] < \infty$ et $\mathbb{E}[Y^2] < \infty$).

Dispersion, variance, standard deviation

Covariance empirique / sample covariance

Étant donné un échantillon appareillé $(x_1, y_1), \cdots, (x_n, y_n)$,

$$\text{cov} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

Example: toss a coin of bias p, with outcome $X \in \{0, 1\}$,

$$\mathbb{E}(X) = p,\ \mathbb{E}(X^2) = p,\ \text{Var}(X) = p - p^2 = p(1-p).$$

$$\text{Var}(aX + b) = a^2 Var(X),\ \forall a, b \in \mathbb{R}, X$$

$$\text{Var}(X_1 + \cdots + X_k) = \text{Var}(X_1) + \cdots + \text{Var}(X_k),\ \text{ if } X_i\text{'s are not correlated}$$

See symmetric random walk, $X_i \in \{-1, +1\}$, $X = X_1 + \cdots + X_n$, then

$$\mathbb{E}(X) = 0,\ \text{Var}(X) = n \text{ and } \text{stdev}(X) = \sqrt{n}$$

Dispersion, variance, standard deviation

Note Même si $\text{Var}(X) = \mathbb{E}\big((X - \mathbb{E}(X))^2\big) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$,

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \neq \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

```
> x = 1:6
> var(x)
[1] 3.5
> mean(x^2)-mean(x)^2
[1] 2.916667
> (6-1)/6*var(x)
[1] 2.916667
```

Dispersion, variance, standard deviation I

Example: The outcome of a (fair) six-sided die has expected value

$$\mathbb{E}[Y] = \sum_{i=1}^{6} \frac{1}{6} i = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = \frac{7}{2}$$

and variance

$$\text{Var}[Y] = \frac{1}{5} \sum_{i=1}^{6} \left(i - \frac{7}{2}\right)^2 = \frac{1}{5} \left[\left(\frac{2-7}{2}\right)^2 + \cdots + \left(\frac{12-7}{2}\right)^2 \right] = \frac{7}{2}$$

```
> x = 1:6
> var(x)
[1] 3.5
```

Dispersion, variance, standard deviation ★★★

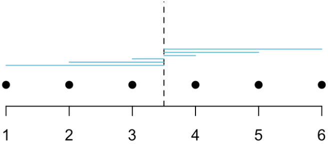
Variance empirique

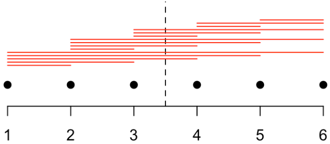
La variance peut s'écrire

$$s^2 = \frac{1}{n(n-1)} \sum_{i<j} (x_i - x_j)^2 = \frac{1}{2n(n-1)} \sum_{i,j=1}^{n} (x_i - x_j)^2 .$$

```
> (D = matrix(1:6,6,6)-matrix(1:6,6,6,byrow = TRUE))
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0   -1   -2   -3   -4   -5
[2,]    1    0   -1   -2   -3   -4
[3,]    2    1    0   -1   -2   -3
[4,]    3    2    1    0   -1   -2
[5,]    4    3    2    1    0   -1
[6,]    5    4    3    2    1    0
> sum(D^2)/(2*5*6)
[1] 3.5
```

Dispersion, variance, standard deviation ★★★





Preuve:

$$\sum_{i,j=1}^{n}(x_i - x_j)^2 = \sum_{i,j=1}^{n}\left(x_i^2 - 2x_ix_j + x_j^2\right)$$

$$= \left(n\sum_{i=1}^{n}x_i^2\right) - 2\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{j=1}^{n}x_j\right) + \left(n\sum_{j=1}^{n}x_j^2\right)$$

or $\displaystyle\sum_{i=1}^{n}x_i^2 = (n-1)s^2 + n\overline{x}^2$ donc

$$\sum_{i,j=1}^{n}(x_i - x_j)^2 = 2n\big((n-1)s^2 + n\overline{x}^2\big) - 2n^2\overline{x}^2 = 2n(n-1)s^2$$

Dispersion, variance, standard deviation

Écart-type (version 2)

Étant donné un échantillon $x_1, \cdots, x_n$, on appelle écart-type

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

```
> x = 1:6
> var(x)
[1] 3.5
> sd(x)
[1] 1.870829
```

Statistique d'ordre et rangs ★★★

Statistique d'ordre

Étant donné un échantillon $\{x_1, \cdots, x_n\}$, on note $\{x_{(1)}, \cdots, x_{(n)}\}$ ou $\{x_{1:n}, \cdots, x_{n:n}\}$ la version ordonnée (dans l'ordre croissant),

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)} \quad \begin{cases} x_{(1)} = \min\{x_1, \cdots, x_n\} \\ x_{(n)} = \max\{x_1, \cdots, x_n\} \end{cases}$$

Ces grandeurs sont liés aux quantiles, et sont parfois utilisés pour les tests, exemple les tests d'indépendance (test du signe), et pour de nombreux tests (Wilcoxon, Mann & Whitney)

Statistique d'ordre et rangs ★★★

Statistique d'ordre

Étant donné un échantillon $\{x_1, \cdots, x_n\}$, le rang de la i-ème observation est r_i tel que $x_i = x_{(r_i)}$,

$$r_i = \sum_{j=1}^{n} \mathbf{1}(x_j \geq x_i)$$

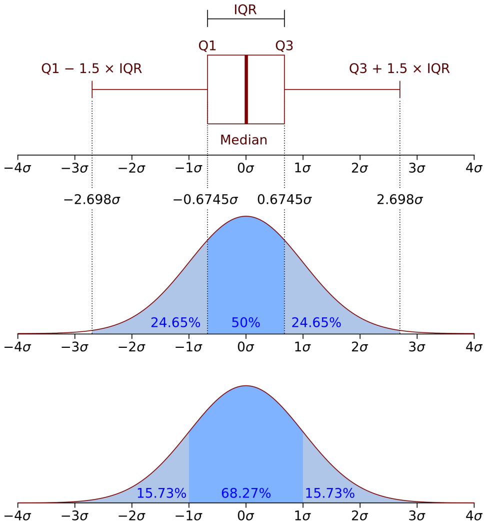
Ces grandeurs sont liés aux quantiles, et sont parfois utilisés pour les tests, exemple les tests d'indépendance (test du signe), et pour de nombreux tests (Wilcoxon, Mann & Whitney)

Intervalle interquartile

Intervalle interquartile (IQR)

$$IQR = Q(3/4) - Q(1/4)$$

cf wikipedia



Coefficient de variation

Si l'écart-type ou l'écart interquartile est une mesure de dispersion de la même unité que x, on peut aussi définir une mesure de dispersion relative,

Coefficient de variation

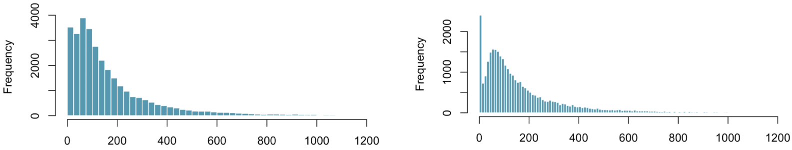
On appelle coefficient de variation le ratio

$$CV(X) = \frac{\sqrt{\text{Var}[X]}}{\mathbb{E}[X]} \text{ et } cv(\boldsymbol{x}) = \frac{s}{\overline{x}}.$$

Statistiques descriptives

31,492 appels au service à la clientèle d'une banque

```
> hist(bankcall$Time[bankcall$Time<1200], breaks=seq
    (0,1200,by=10))
```




```
> mean(bankcall$Time/60)
[1] 3.143204
```

soit 3 minutes et 8 secondes pour la moyenne

```
> median(bankcall$Time/60)
[1] 1.916667
```

soit 1 minute et 55 secondes pour la médiane (115 sec.)

Statistiques descriptives

La moyenne tronquée (trimmed mean en anglais) à α (5% ou 10%) est la moyenne du jeu de données obtenu en supprimant une proportion α des plus petites valeurs et α plus grandes valeurs :

```
> mean(bankcall$Time/60, trim=.1)
[1] 2.357288
```

```
> mean(bankcall$Time/60)
[1] 3.143204
```

```
> quantile(bankcall$Time)
   0%    25%    50%    75%   100%
    1     57    115    225  28739
```

Moyenne, Variance, Quantiles

Ajouter une constante

Soit X une variable aléatoire, et $Y = a + X$,

$$\mathbb{E}[Y] = a + \mathbb{E}[X], \ \text{Var}[Y] = \text{Var}[X] \text{ et } Q_Y(p) = a + Q_X(p)$$

Multiplier par une constante

Soit X une variable aléatoire, et $Y = b \cdot X$,

$$\mathbb{E}[Y] = b \cdot \mathbb{E}[X], \ \text{Var}[Y] = b^2 \cdot \text{Var}[X] \text{ et } Q_Y(p) = b \cdot Q_X(p)$$

Example: changement d'unité

Calculs Formels

Exemple: $X \sim \mathcal{N}(1,3^2)$ et $Y \sim \mathcal{E}(2)$, indépendantes. Que vaut $E_1 = \mathbb{E}[(X^2-1)Y]$? $E_1 = \dfrac{9}{2}$

Exemple: $\mathbb{P}[X \leq x] = \dfrac{x^2-2x+2}{2}$ sur $[1,2]$, 0 avant et 1 après. que vaut $E_2 = \mathbb{E}[X]$?

Exemple: $\mathbb{P}[X \leq x] = \dfrac{x}{8}$ sur $[0,2)$, $\dfrac{x^2}{16}$ sur $[2,4)$, 0 avant et 1 après. Que vaut $E_3 = \text{Var}[X-1]+1$? $E_3 = \dfrac{311}{144}$

Exemple:

$$f_X(x) = \begin{cases} 2(3-2x)/5 & \text{pour } 0 \leq x \leq 1 \\ 2(2-x)/5 & \text{pour } 1 \leq x \leq 2 \\ 0 & \text{sinon} \end{cases}$$

Que vaut m_1 la médiane de X? $m_1 = \frac{1}{2}$

Calculs Numériques

Exemple: $X \sim \mathcal{P}(10)$, $p_1 = \mathbb{P}[X > 20]$? $p_1 \approx 0.001588$
Exemple: $X \sim \mathcal{B}(.1, 50)$, $p_2 = \mathbb{P}\big[X \leq \mathbb{E}(X)\big]$? $p_2 \approx 0.6161$
Exemple: $X \sim \mathcal{N}(10, 5)$, $p_3 = \mathbb{P}\big[X \leq 0\big]$? $p_3 \approx 0.02275$

Exemple: $X \sim \mathcal{N}(10, 5)$, q_1 tel que $\mathbb{P}\big[X > q_1\big] = 20\%$?
$q_1 \approx 14.208$
Exemple: $X \sim \mathcal{N}(0, 1)$ et $Y \sim \mathcal{N}(0, 1)$, indépendantes, q_3 tel que
$\mathbb{P}\left[\frac{X_1}{X_2^2} > q_3\right] = 10\%$? $q_3 \approx 10.4079$
Exemple: $U_1, \cdots, U_4 \sim \mathcal{U}_{[0,1]}$, indépendantes, q_4 tel que
$\mathbb{P}\left[X_1 + X_2 + X_3 + X_4 \leq q_4\right] = 10\%$? $q_4 \approx 1.2465$