Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

Rappels #4.1 (statistique, simulations & bootstrap)

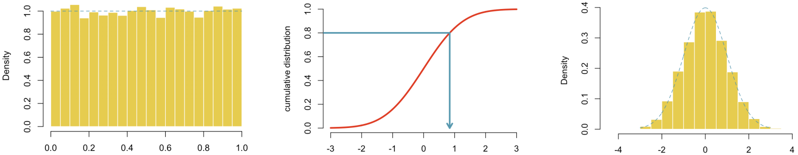
A probabilistic result

If F is a cdf, and if $U \sim \mathcal{U}([0,1])$, $X = F^{-1}(U)$ has cdf F
(see inverse method sampling)

Proof: Let $x \in \mathbb{R}$, $\mathbb{P}[X \leq x]$ is equal to

$$\mathbb{P}[F^{-1}(U) \leq x] = \mathbb{P}[F\big(F^{-1}(U)\big) \leq F\big(x\big)] = \mathbb{P}[U \leq F\big(x\big)] = F(x)$$

where $F^{-1}(u) = \inf \{x \mid F(x) \geq u\}$ for $u \in (0,1)$.



A probabilistic result

```
> U = runif(100)
  [1] 0.26 0.35 0.31 0.76 0.52 0.06 0.03 0.23 0.67 0.14
 [11] 0.17 0.13 0.58 0.93 0.32 0.11 0.53 0.13 0.09 0.19
 [21] 0.32 0.37 0.91 0.47 0.28 0.38 0.88 0.98 0.49 0.84
 [31] 0.51 0.63 0.14 0.60 0.79 0.17 0.37 0.33 0.46 0.72
 [41] 0.92 0.39 0.42 0.48 0.70 0.30 0.05 0.51 0.38 0.27
 [51] 0.51 0.69 0.21 0.11 0.17 0.19 0.14 0.68 0.99 0.50
 [61] 0.26 0.69 0.43 0.25 0.06 0.26 0.32 0.10 0.18 0.08
 [71] 0.05 0.55 0.13 0.50 0.75 0.18 0.15 0.12 0.81 0.35
```

```
> Q(U)
  [1]  1.04 -0.48  0.81 -0.86 -0.33  0.74  0.92  0.38
  [9] -0.80  0.95 -0.76  0.22  0.44  0.77  0.25 -1.45
 [17]  0.10 -0.12 -1.87  0.68  0.73 -1.06 -0.19 -0.19
 [25] -1.10 -0.48  1.09  1.11  0.06  0.04  0.15  0.08
 [33] -0.45 -1.29  0.48 -0.33  0.95  0.25  0.80  1.58
 [41]  0.31 -1.51  1.57  0.84  0.07  0.01 -0.96  0.56
 [49] -0.66  0.49  0.46 -1.57  0.00 -0.29  1.89  0.60
 [57]  0.34  0.43  1.01  0.31 -0.20 -0.19 -0.07 -0.07
 [65] -0.04  1.31 -0.35 -0.37 -0.35 -2.26  1.47 -1.17
```

Notations & Results

Given a sample $\{x_1, \cdots, x_n\}$ i.i.d. from F,

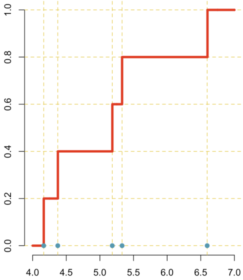
$$F(x0) = \mathbb{P}[X \leq x],$$

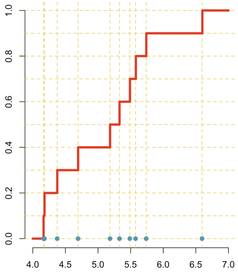
the empirical cumulative distribution function is

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(x_i \leq x),\ x \in \mathbb{R}$$

Glivenko-Cantelli: $\widehat{F}_n \to F$ as $n \to \infty$,
or more precisely, almost surely

$$\|\widehat{F}_n - F\|_\infty = \sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F(x)| \longrightarrow 0$$





A probabilistic result

The inverse method with $\widehat{F}_n$ simply means resampling within $\{x_1, \cdots, x_n\}$ with equal probabilities $1/n$ (or *with replacement*)

```
> x
[1] 4.164 4.374 5.184 5.330 6.595
> Qemp(U)
 [1] 6.60 6.60 6.60 5.33 4.37 5.33 5.33 4.16 6.60 5.33
[11] 4.37 4.37 4.37 6.60 5.33 5.18 5.33 5.18 6.60 5.18
[21] 5.18 4.37 6.60 4.37 4.16 6.60 4.16 6.60 5.33 4.16
[31] 4.16 6.60 4.37 4.37 5.33 5.18 5.18 5.18 5.33 5.33
[41] 4.37 5.18 5.33 5.18 4.37 5.18 5.18 5.18 5.33 5.18
[51] 5.33 4.37 4.37 4.16 5.18 5.18 5.18 5.18 4.16 5.18
[61] 4.37 4.16 4.16 4.16 6.60 4.37 4.37 5.33 5.18 4.16
[71] 5.33 4.16 6.60 5.18 4.16 4.16 5.18 4.16 5.18 4.16
```

called bootstrapping

Bootstrap

Real World:

- distribution F
- data $\{x_1, \cdots, x_n\}$, i.i.d., F
- empirical distribution $\widehat{F}_n$
- parameter $\theta = t(F)$
- estimate $\widehat{\theta}_n = t(\widehat{F}_n)$
- error $\widehat{\theta}_n - \theta$
- standardized error $\dfrac{\widehat{\theta}_n - \theta}{s(\widehat{F}_n)}$

Bootstrap World ($\star$):

- distribution $\widehat{F}_n$
- data $\{x_1^\star, \cdots, x_n^\star\}$, i.i.d., $\widehat{F}_n$
- empirical distribution $\widehat{F}_n^\star$
- parameter $\widehat{\theta}_n = t(\widehat{F}_n)$
- estimate $\widehat{\theta}_n^\star = t(\widehat{F}_n^\star)$
- error $\widehat{\theta}_n^\star - \widehat{\theta}_n$
- standardized error $\dfrac{\widehat{\theta}_n^\star - \widehat{\theta}_n}{s(\widehat{F}_n^\star)}$

The sampling distribution of $\widehat{\theta}_n$ depends on (unknown) F
Use $\widehat{F}_n$ as a proxy for F: we cannot resample from F, but we can from $\widehat{F}_n$

Example: mean, $\theta = t(F) = \int x dF(x)$

$$\widehat{\theta}_n = t(\widehat{F}_n) = \int x d\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: variance, $\theta = t(F) = \int x^2 dF(x) - \left(\int x dF(x)\right)^2$

$$\widehat{\theta}_n = t(\widehat{F}_n) = \int x^2 d\widehat{F}_n(x) - \left(\int x d\widehat{F}_n(x)\right) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

Bootstrap& Confidence Intervals

Let H_n be the true distribution of $\widehat{\theta}_n = t(\widehat{F}_n)$
Let $H_n^\star$ be the true distribution of $\widehat{\theta}_n^\star = t(\widehat{F}_n^\star)$

Importance application of bootstrap: construct confidence intervals
percentile method: let $\widehat{\theta}_n^{\star 1}, \cdots, \widehat{\theta}_n^{\star B}$ be the bootstrap sample of estimators). Let $\widehat{H}_n^\star$ denote its empirical distribution,

$$\widehat{H}_n^\star(\theta) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(\widehat{\theta}_n^{\star b} \leq \theta)$$

and consider $\left[\widehat{H}_n^{\star -1}(\alpha/2); \widehat{H}_n^{\star -1}(1-\alpha/2)\right]$

Bootstrap & Confidence Intervals

Studentized method: Let $\widehat{\theta}_n$ be an estimate of θ, and let $\widehat{\sigma}_n$ be an estimate of standard deviation of $\widehat{\theta}_n$.

Let $t_n = \dfrac{\widehat{\theta}_n - \theta}{\widehat{\sigma}_n}$ denote the t-statistic. Its bootstrap counterpart is

$$t_n^\star = \frac{\widehat{\theta}_n^\star - \widehat{\theta}_n}{\widehat{\sigma}_n^\star}$$

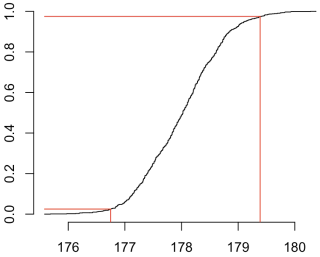
Then the confidence interval for θ is

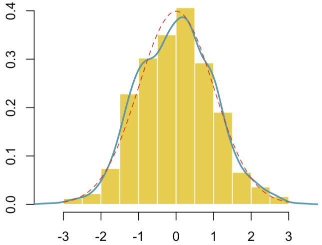
$$\left[\widehat{\theta}_n + u^\star_{\alpha/2}\widehat{\sigma}_n ; \widehat{\theta}_n + u^\star_{1-\alpha/2}\widehat{\sigma}_n\right]$$

where $u_p^\star$ is the p-quantile of $t_n^\star$.

Bootstrap

```
> M = rep(NA,999)
> for(b in 1:999){
+   i = sample(1:length(X),length(X),
+   replace=TRUE)
+   M[b] = mean(X[i])
+ }
> quantile(M,c(.025,.975))
     2.5%     97.5%
176.7494 179.3875
> s = sd(M)
> T = (M-mean(X))/s
> u = quantile(T,c(.025,.975))
> mean(X) + u*s
     2.5%     97.5%
176.7440 179.3935
```





Back on theoretical results on $\widehat{\theta}$ (MLE)

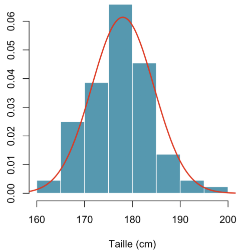
Suppose that the height of male students has a Gaussian distribution, $\mathcal{N}(\theta, 6.5^2)$

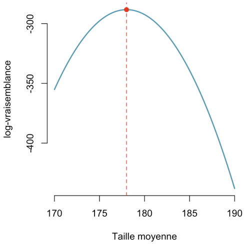
```
> X = Davis$height[Davis$sex=="M"]
> logL = function(m) sum(log(dnorm(X,
    mean=m,sd=6.5)))
> optim(par=180, fn=function(x) -logL(x)
    )
$par
[1] 178.0093

$value
[1] 288.2951
```

We've seen that

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta))$$

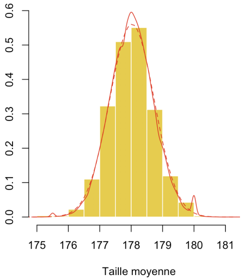




Back on theoretical results on $\widehat{\theta}$ (MLE)

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$$

```
> P=rep(NA,99999)
> for(b in 1:99999){
  Xb = sample(X,size=length(X),replace=
    TRUE)
  logL = function(m) -sum(log(dnorm(Xb,
    mean=m,sd=6.5)))
  P[b]=optim(par=180, logL)$par
}
> hist(P,probability = TRUE)
> lines(density(P))
```



We've seen here that

$$(\widehat{\boldsymbol{\theta}}_n - \star) \overset{\mathcal{L}}{\approx} \mathcal{N}(0, \star)$$

Back on theoretical results on $\widehat{\theta}$ (MLE)

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta))$$

Let us simule data from a $\mathcal{N}(180, 6.5^2)$ distribution

```
> P=rep(NA,99999)
> for(b in 1:99999){
  Xb = rnorm(length(X),180,6.5)
  logL = function(m) -sum(log(dnorm(Xb,
    mean=m,sd=6.5)))
  P[b]=optim(par=180, logL)$par
}
> hist(P,probability = TRUE)
> lines(density(P))
> var(P)
[1] 0.481649
> var(X)/length(X)
[1] 0.4713935
```