



STT 1000 - STATISTIQUES

ARTHUR CHARPENTIER



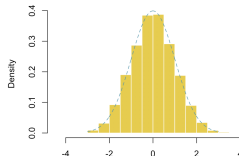
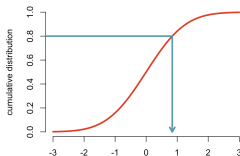
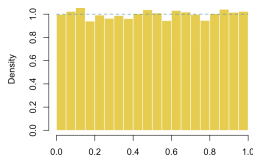
A probabilistic result

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

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

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



A probabilistic result

```
1 > U = runif(100)
2 [1] 0.26 0.35 0.31 0.76 0.52 0.06 0.03 0.23 0.67 0.14
3 [11] 0.17 0.13 0.58 0.93 0.32 0.11 0.53 0.13 0.09 0.19
4 [21] 0.32 0.37 0.91 0.47 0.28 0.38 0.88 0.98 0.49 0.84
5 [31] 0.51 0.63 0.14 0.60 0.79 0.17 0.37 0.33 0.46 0.72
6 [41] 0.92 0.39 0.42 0.48 0.70 0.30 0.05 0.51 0.38 0.27
7 [51] 0.51 0.69 0.21 0.11 0.17 0.19 0.14 0.68 0.99 0.50
8 [61] 0.26 0.69 0.43 0.25 0.06 0.26 0.32 0.10 0.18 0.08
9 [71] 0.05 0.55 0.13 0.50 0.75 0.18 0.15 0.12 0.81 0.35
```

```
1 > Q(U)
2 [1] 1.04 -0.48 0.81 -0.86 -0.33 0.74 0.92 0.38
3 [9] -0.80 0.95 -0.76 0.22 0.44 0.77 0.25 -1.45
4 [17] 0.10 -0.12 -1.87 0.68 0.73 -1.06 -0.19 -0.19
5 [25] -1.10 -0.48 1.09 1.11 0.06 0.04 0.15 0.08
6 [33] -0.45 -1.29 0.48 -0.33 0.95 0.25 0.80 1.58
7 [41] 0.31 -1.51 1.57 0.84 0.07 0.01 -0.96 0.56
8 [49] -0.66 0.49 0.46 -1.57 0.00 -0.29 1.89 0.60
9 [57] 0.34 0.43 1.01 0.31 -0.20 -0.19 -0.07 -0.07
10 [65] -0.04 1.31 -0.35 -0.37 -0.35 -2.26 1.47 -1.17
```

Notations & Results

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

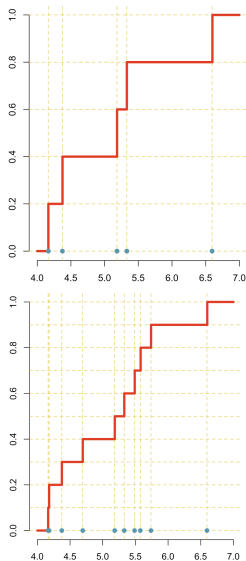
$$F(x) = \mathbb{P}[X \leq x],$$

the empirical cumulative distribution function is

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

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

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



A probabilistic result

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

```
1 > x
2 [1] 4.164 4.374 5.184 5.330 6.595
3 > Qemp(U)
4 [1] 6.60 6.60 6.60 5.33 4.37 5.33 5.33 4.16 6.60 5.33
5 [11] 4.37 4.37 4.37 6.60 5.33 5.18 5.33 5.18 6.60 5.18
6 [21] 5.18 4.37 6.60 4.37 4.16 6.60 4.16 6.60 5.33 4.16
7 [31] 4.16 6.60 4.37 4.37 5.33 5.18 5.18 5.18 5.33 5.33
8 [41] 4.37 5.18 5.33 5.18 4.37 5.18 5.18 5.18 5.33 5.18
9 [51] 5.33 4.37 4.37 4.16 5.18 5.18 5.18 5.18 4.16 5.18
10 [61] 4.37 4.16 4.16 4.16 6.60 4.37 4.37 5.33 5.18 4.16
11 [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 $\{x_1, \dots, x_n\}$, i.i.d.,
 F
- empirical distribution \hat{F}_n
- parameter $\theta = t(F)$
- estimate $\hat{\theta}_n = t(\hat{F}_n)$
- error $\hat{\theta}_n - \theta$
- standardized error $\frac{\hat{\theta}_n - \theta}{s(\hat{F}_n)}$

Bootstrap World (\star):

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

The sampling distribution of $\hat{\theta}_n$ depends on (unknown) F

Use \hat{F}_n as a proxy for F : we cannot resample from F , but we can from \hat{F}_n

Bootstrap

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

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

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

$$\hat{\theta}_n = t(\hat{F}_n) =$$

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

Bootstrap & Confidence Intervals

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

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

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

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

and consider $[\hat{H}_n^{*-1}(\alpha/2); \hat{H}_n^{*-1}(1 - \alpha/2)]$

Bootstrap & Confidence Intervals

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

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

$$t_n^* = \frac{\hat{\theta}_n^* - \hat{\theta}_n}{\hat{\sigma}_n^*}$$

Then the confidence interval for θ is

$$\left[\hat{\theta}_n + u_{\alpha/2}^* \hat{\sigma}_n; \hat{\theta}_n + u_{1-\alpha/2}^* \hat{\sigma}_n \right]$$

where u_p^* is the p -quantile of t_n^* .

Bootstrap

```
1 > M = rep(NA,999)
2 > for(b in 1:999){
3 +   i = sample(1:length(X),length(X),
4 +   replace=TRUE)
5 +   M[b] = mean(X[i])
6 + }
7 > quantile(M,c(.025,.975))
8     2.5%     97.5%
9 176.7494 179.3875
10 > s = sd(M)
11 > T = (M-mean(X))/s
12 > u = quantile(T,c(.025,.975))
13 > mean(X) + u*s
14     2.5%     97.5%
15 176.7440 179.3935
```

figs/boot-ic-percent-1

figs/boot-ic-percent-2

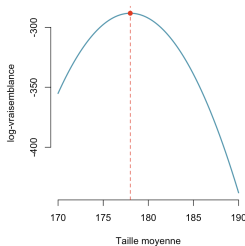
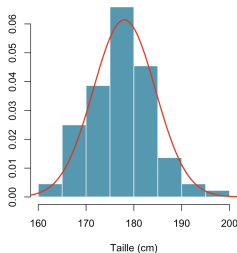
Back on theoretical results on $\hat{\theta}$ (MLE)

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

```
1 > X = Davis$height[Davis$sex=="M"]
2 > logL = function(m) sum(log(dnorm(X,
   mean=m, sd=6.5)))
3 > optim(par=180, fn=function(x) -logL(x)
   )
4 $par
5 [1] 178.0093
6
7 $value
8 [1] 288.2951
```

We've seen that

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

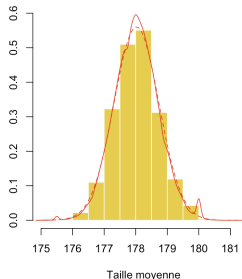


Back on theoretical results on $\hat{\theta}$ (MLE)

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

```
1 > P=rep(NA,99999)
2 > for(b in 1:99999){
3   Xb = sample(X,size=length(X),replace=
4     TRUE)
5   logL = function(m) -sum(log(dnorm(Xb,
6     mean=m,sd=6.5)))
7   P[b]=optim(par=180, logL)$par
8 }
```

```
1 > hist(P,probability = TRUE)
2 > lines(density(P))
```



We've seen here that

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

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

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

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

```
1 > P=rep(NA,99999)
2 > for(b in 1:99999){
3   Xb = rnorm(length(X),180,6.5)
4   logL = function(m) -sum(log(dnorm(Xb,
5     mean=m,sd=6.5)))
6   P[b]=optim(par=180, logL)$par
7 }
8 > hist(P,probability = TRUE)
9 > lines(density(P))
10 > var(P)
11 [1] 0.481649
12 > var(X)/length(X)
13 [1] 0.4713935
```

