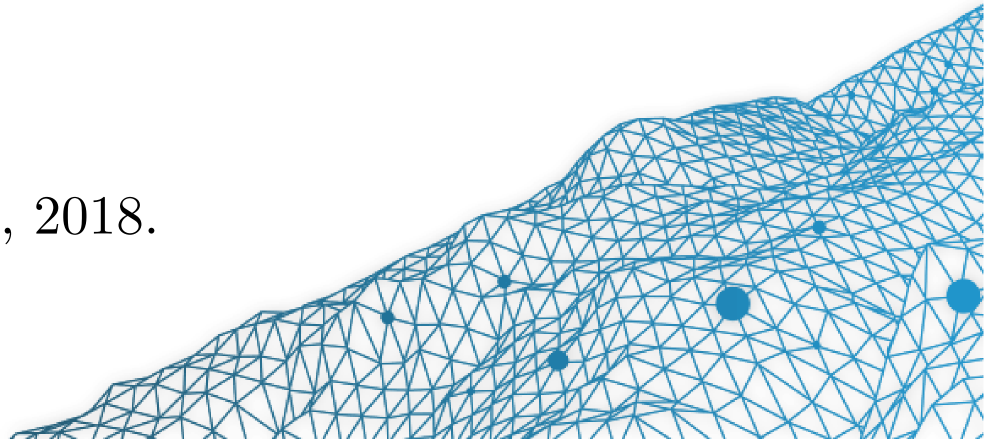
Big Data for Economics # 2

A. Charpentier (Université de Rennes 1)

https://github.com/freakonometrics/ub

UB School of Economics Summer School, 2018.



#2 Simulation Based Techniques & Bootstrap

Historical References

Permutation methods go back to Fisher (1935, The Design of Experiments and Pitman (1937) Significance tests which may be applied to samples from any population)

(there are $n!$ distinct permutations)

Jackknife was introduced in Quenouille (1949, Approximate tests of correlation in time series), popularized by Tukey (1958, Bias and confidence in not quite large samples)

Bootstrapping started with Monte Carlo algorithms in the 40's, see e.g. Simon & Burstein (1969, Basic Research Methods in Social Science)

Efron (1979, Bootstrap methods: Another look at the jackknife) defined a resampling procedure that was coined as "bootstrap".

(there are n^n possible distinct ordered bootstrap samples)

References

Motivation

Bertrand, M., Duflo, E. & Mullainathan, 2004. Should we trust difference-in-difference estimators?. QJE.

References

Davison, A.C. & Hinkley, D.V. 1997 Bootstrap Methods and Their Application. CUP.

Efron B. & Tibshirani, R.J. An Introduction to the Bootstrap. CRC Press.

Horowitz, J.L. 1998 The Bootstrap, Handbook of Econometrics, North-Holland.

MacKinnon, J. 2007 Bootstrap Hypothesis Testing, Working Paper.

Complex Computations? Use "*simulations*"...

Consider a sample $\{y_1, \cdots, y_n\}$. The natural estimator of the variance is

$$\widehat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$$

What is the variance of that estimator ? If y_i's are obtained from i.i.d. normal random variables, then $\text{Var}[\widehat{\sigma}^2] = \dfrac{2\sigma^4}{n-1}$, so the standard error of $\widehat{\sigma}^2$ can be estimated as

$$\widehat{\text{se}}[\widehat{\sigma}^2] = \frac{\sqrt{2}\widehat{\sigma}^2}{\sqrt{n-1}}$$

What if the sample is not normally distributed ?

Preliminaries: Generating Randomness

TABLE OF RANDOM DIGITS 1

00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527

Source A Million Random Digits with 100,000 Normal Deviates, RAND, 1955.

Preliminaries: Generating Randomness

Here random means a sequence of numbers do not exhibit any discernible pattern, i.e. successively generated numbers can not be predicted.

A random sequence is a vague notion... in which each term is unpredictable to the uninitiated and whose digits pass a certain number of tests traditional with statisticians... Derrick Lehmer, quoted in Knuth (1997)

The goal of Pseudo-Random Numbers Generators is to produce a sequence of numbers in $[0, 1]$ that imitates ideal properties of random number.

```
> runif(30)
 [1] 0.3087420 0.4481307 0.0308382 0.4235758 0.7713879 0.8329476
 [7] 0.4644714 0.0763505 0.8601878 0.2334159 0.0861886 0.4764753
[13] 0.9504273 0.8466378 0.2179143 0.6619298 0.8372218 0.4521744
[19] 0.7981926 0.3925203 0.7220769 0.3899142 0.5675318 0.4224018
[25] 0.3309934 0.6504410 0.4680358 0.7361024 0.1768224 0.8252457
```

Linear Congruential Method

Produce a sequence of integers $X_1, X_2, \cdots$ between 0 and $m-1$ following a recursive relationship $X_{i+1} = (aX_i + b)$ modulo m, and set $U_i = X_i/m$.

E.g. Start with $X_0 = 77$, $a = 13$, $b = 43$ and $m = 100$. Then the sequence is

$$\{77, 52, 27, 2, 77, 52, 27, 2, 77, 52, 27, 2, 77, \cdots\}$$

Problem: not all values in $\{0, \cdots, m-1\}$ are obtained, and there is a cycle here.

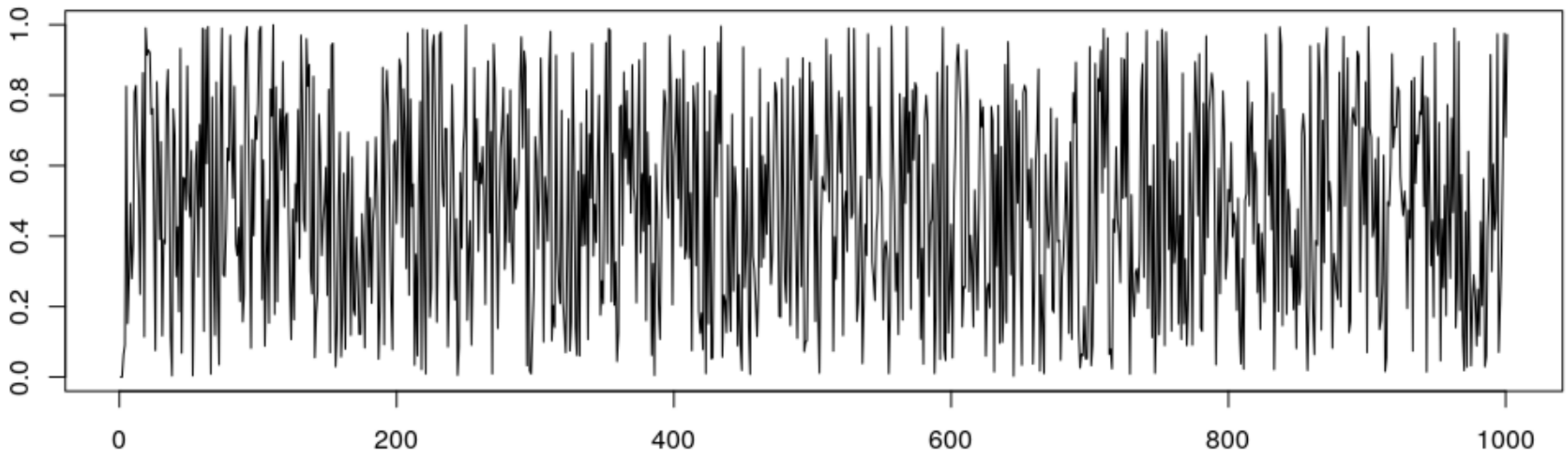
Solution: use (very) large values for m and choose properly a and b.

E.g. $m = 2^{32} - 1$, $a = 16807$ $(= 7^5)$ and $b = 0$ (used in Matlab).

Linear Congruential Method

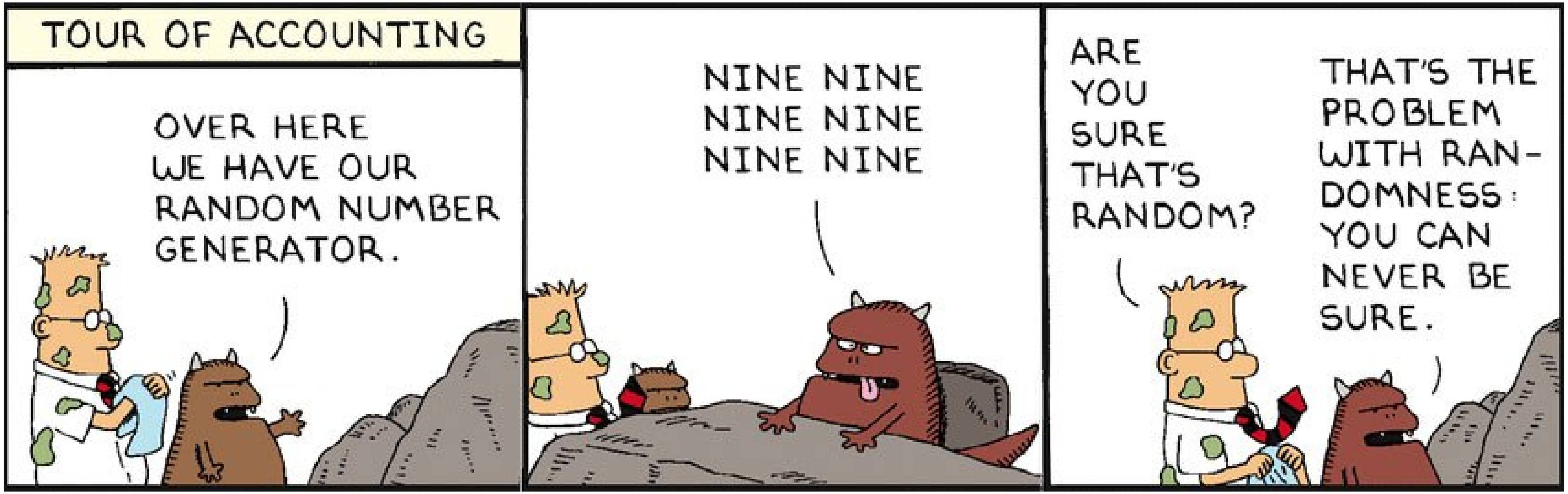
If we start with $X_0 = 77$, we get for $U_{100}, U_{101}, \cdots$

$$\{\cdots, 0.9814, 0.9944, 0.2205, 0.6155, 0.0881, 0.3152, 0.5028, 0.1531, 0.8171, 0.7405, \cdots\}$$



See L'Ecuyer (2017) for an historical perspective.

Randomness?



Source Dibert, 2001.

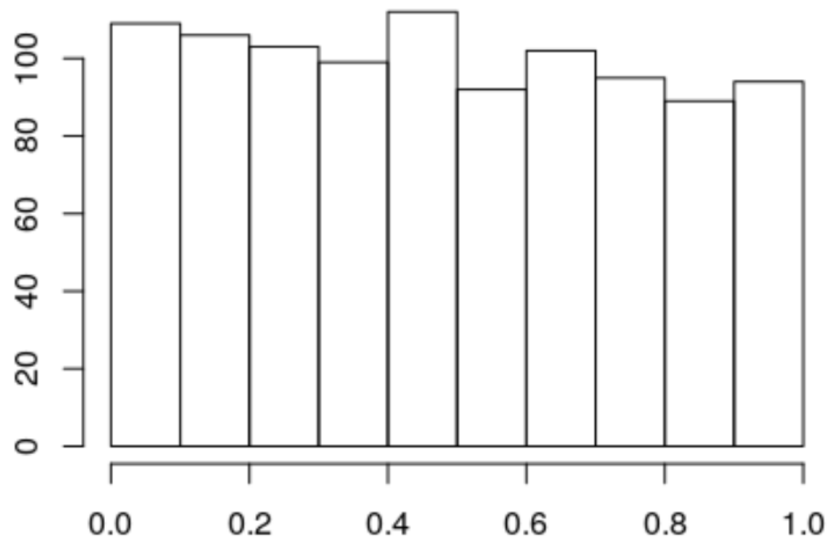
Randomness?

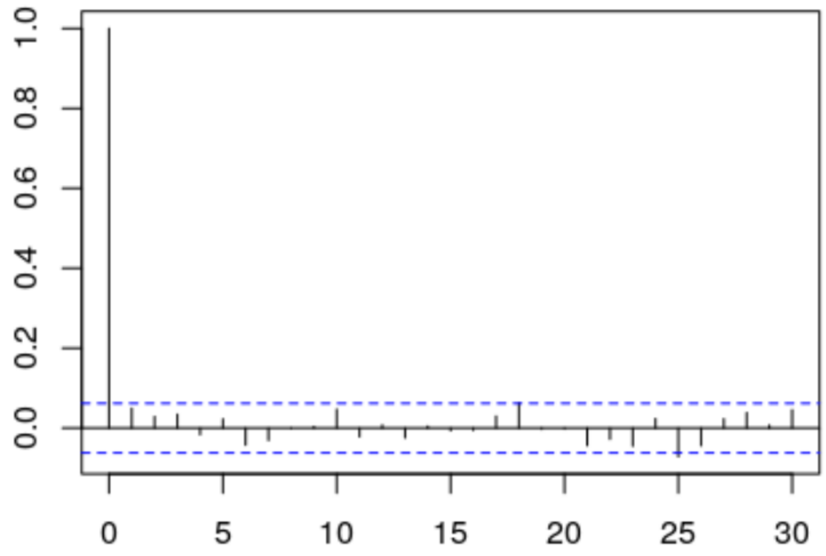
Heuristically,

1. calls should provide a uniform sample, $\displaystyle\lim_{n\rightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{u_i\in(a,b)}=b-a$ with $b>a$,

2. calls should be independent, $\displaystyle\lim_{n\rightarrow\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{u_i\in(a,b),u_{i+k}\in(c,d)}=(b-a)(d-c)$

$\forall k\in\mathbb{N}$, and $b>a$, $d>c$.





Monte Carlo: from $\mathcal{U}_{[0,1]}$ to any distribution

Recall that the cumulative distribution function of Y is $F : \mathbb{R} \to [0, 1]$, $F(y) = \mathbb{P}[Y \leq y]$.

Since F is an increasing function, define its (pseudo-)inverse $Q : (0, 1) \to \mathbb{R}$ as

$$Q(u) = \inf \left\{ y \in \mathbb{R} : F(y) > u \right\}$$

Proposition If $U \sim \mathcal{U}_{[0,1]}$, then $Q(U) \sim F$.

Monte Carlo

From the law of large numbers, if $U_1, U_2, \cdots$ is a sequence of i.i.d random variables, uniformly distributed on $[0,1]$, and some mapping $h:[0,1]\rightarrow\mathbb{R}$,

$$\frac{1}{n}\sum_{i=1}^n h(U_i)\xrightarrow{\text{a.s.}}\mu=\int_{[0,1]}h(u)\,\mathrm{d}u=\mathbb{E}[h(U)],\text{ as }\ n\rightarrow\infty$$

and from the central limit theorem

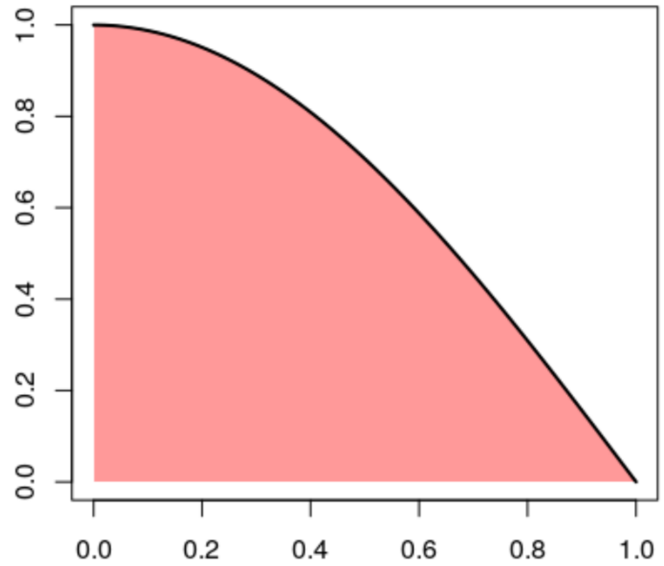
$$\sqrt{n}\left(\left(\frac{1}{n}\sum_{i=1}^n h(U_i)\right)-\mu\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\sigma^2\right)$$

where $\sigma^2=\text{Var}[h(U)]$, and $U\sim\mathcal{U}_{[0,1]}$.

Monte Carlo

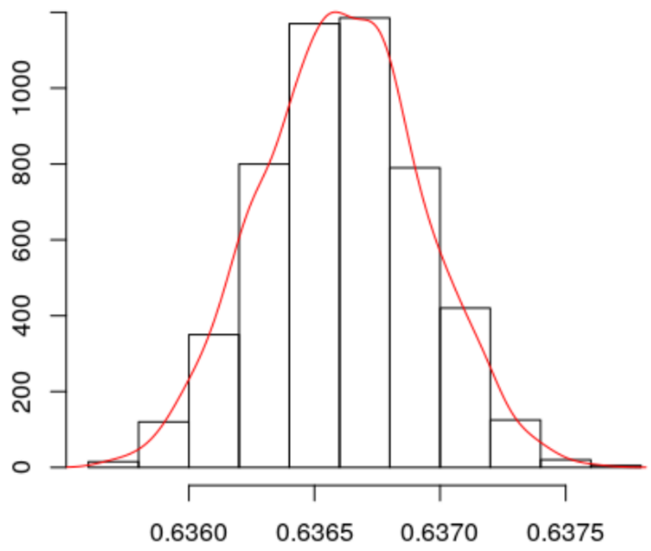
Consider $h(u) = \cos(\pi u/2)$,

```
> h=function(u) cos(u*pi/2)
> integrate(h,0,1)
0.6366198 with absolute error<7.1e-15
> mean(h(runif(1e6)))
[1] 0.6363378
```



We can actually repeat that a thousand time

```
> M=rep(NA,1000)
> for(i in 1:1000) M[i]=mean(h(runif(1e6)))
> mean(M)
[1] 0.6366087
> sd(M)
[1] 0.000317656
```



Monte Carlo Techniques to Compute Integrals

Monte Carlo is a very general technique, that can be used to compute any integral.

Let $X \sim \mathcal{C}auchy$ what is $\mathbb{P}[X > 2]$. Observe that

$$\mathbb{P}[X > 2] = \int_2^\infty \frac{dx}{\pi(1+x^2)} \qquad (\sim 0.15)$$

since $f(x) = \dfrac{1}{\pi(1+x^2)}$ and $Q(u) = F^{-1}(u) = \tan\left(\pi\left[u - \frac{1}{2}\right]\right)$.

Crude Monte Carlo: use the law of large numbers

$$\widehat{p}_1 = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Q(u_i) > 2)$$

where u_i are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_1] \sim \frac{0.127}{n}$.

Crude Monte Carlo (with symmetry): $\mathbb{P}[X > 2] = \mathbb{P}[|X| > 2]/2$ and use the law of large numbers

$$\widehat{p}_2 = \frac{1}{2n}\sum_{i=1}^{n}\mathbf{1}(|Q(u_i)| > 2)$$

where u_i are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_2] \sim \frac{0.052}{n}$.

Using integral symmetries :

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \frac{1}{2} - \int_0^2 \frac{dx}{\pi(1+x^2)}$$

where the later integral is $\mathbb{E}[h(2U)]$ where $h(x) = \dfrac{2}{\pi(1+x^2)}$.

From the law of large numbers

$$\widehat{p}_3 = \frac{1}{2} - \frac{1}{n}\sum_{i=1}^{n} h(2u_i)$$

where u_i are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_3] \sim \frac{0.0285}{n}$.

Using integral transformations :

$$\int_2^\infty \frac{dx}{\pi(1+x^2)} = \int_0^{1/2} \frac{y^{-2}dy}{\pi(1-y^{-2})}$$

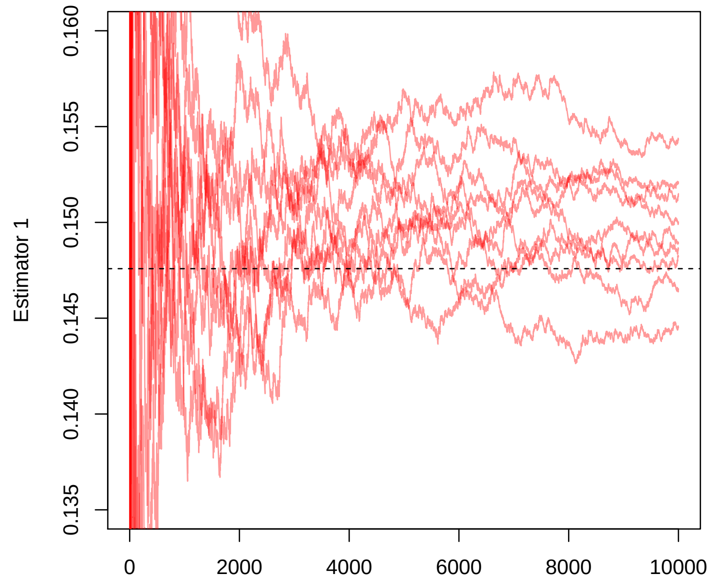
which is $\mathbb{E}[h(U/2)]$ where $h(x) = \dfrac{1}{2\pi(1+x^2)}$.

From the law of large numbers

$$\widehat{p}_4 = \frac{1}{4n}\sum_{i=1}^n h(u_i/2)$$

where u_i are obtained from i.id. $\mathcal{U}([0,1])$ variables.

Observe that $\text{Var}[\widehat{p}_4] \sim \frac{0.0009}{n}$.



Kolmogorov-Smirnov Test and Monte Carlo

Kolmogorov-Smirnov test, $H_0 : F = F_0$ (against $H_1 : F \neq F_0$). The test statistic for a given cdf F_0 is

$$D_n = \sup_x \left\{\left|\widehat{F}_n(x) - F(x)\right|\right\}$$

One can prove that under H_0, $\sqrt{n}D_n \xrightarrow{\mathcal{L}} \sup_t |B_{F(t)}|$, as $n \to \infty$, where (B_t) is the Brownian bridge on $[0, 1]$.

Consider the height of 200 students.

```
> Davis = read.table("http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/Davis.txt")
> Davis[12,c(2,3)]=Davis[12,c(3,2)]
> Y = Davis$height
> mean(Y)
[1] 170.565
> sd(Y)
[1] 8.932228
```

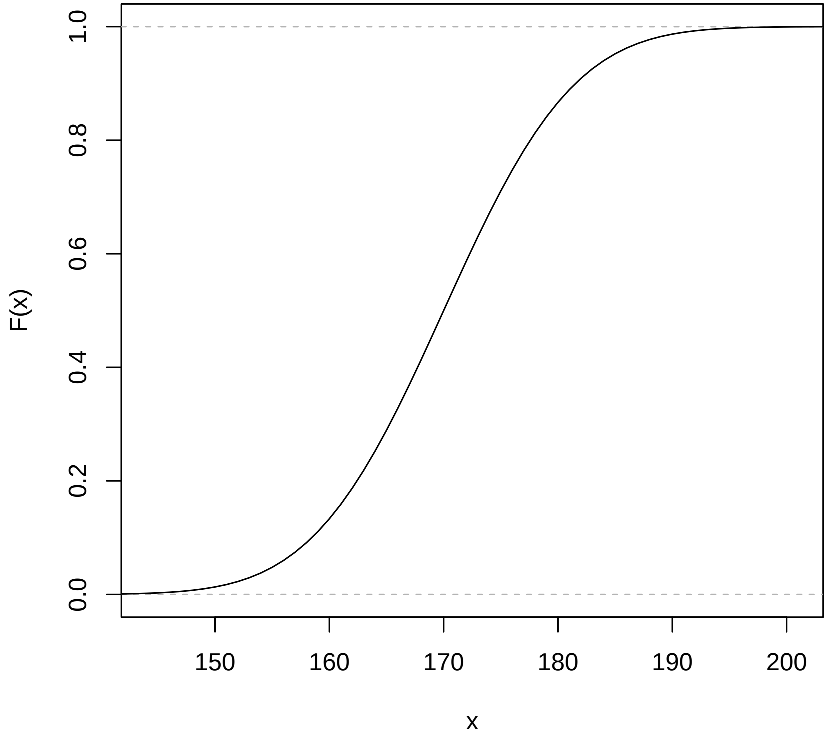
Kolmogorov-Smirnov Test and Monte Carlo

Let us test $F = \mathcal{N}(170, 9^2)$.

```
> y0 = pnorm(140:205,170,9)
> for(s in 1:200){
+    X = rnorm(length(Y),170,9)
+    y = Vectorize(ecdf(X))(140:205)
+    lines(140:205,y)
+    D[s] = max(y-y0)
+ }
```

while for $\widehat{F}_n$,

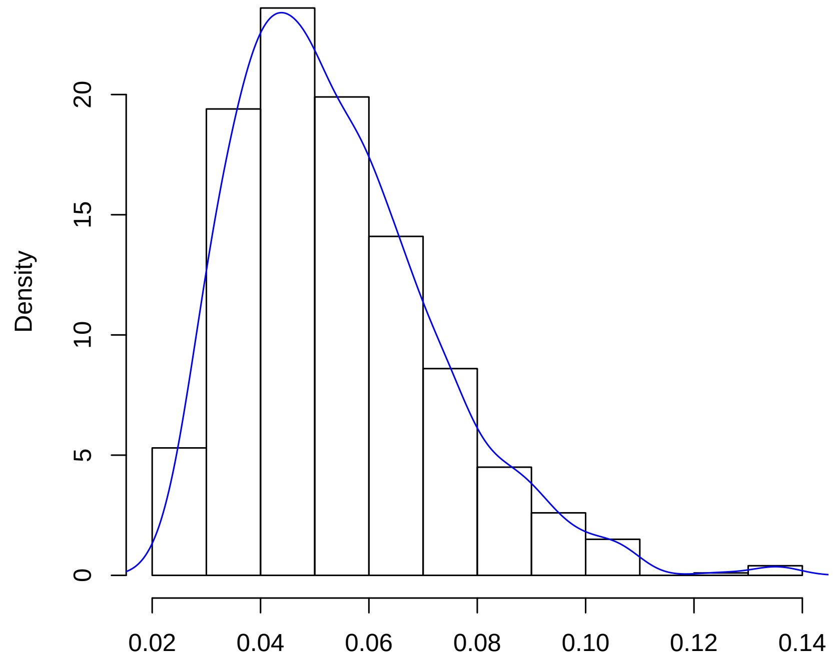
```
> lines(140:205, Vectorize(ecdf(Y))
   (140:205)),col="red")
```



Kolmogorov-Smirnov Test and Monte Carlo

The empirical distribution of D is obtained using

```
> hist(D,probability = TRUE )
> lines(density(D),col="blue")
```



Here

```
> (demp = max(abs(Vectorize(ecdf(Y))
    (140:205)-y0)))
[1] 0.05163936
> mean(D>demp)
[1] 0.2459
```

```
> ks.test(Y, "pnorm",170,9)
D = 0.062969, p-value = 0.406
```

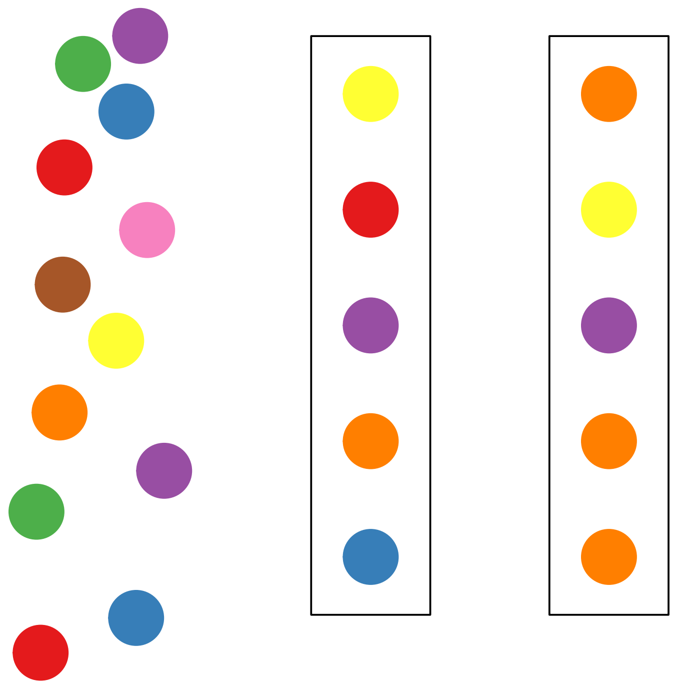
Bootstrap Techniques (in one slide)

Bootstrapping is an asymptotic refinement based on computer based simulations.

Underlying properties: we know when it might work, or not

Idea : $\{(y_i, \boldsymbol{x}_i)\}$ is obtained from a stochastic model under $\mathbb{P}$

We want to generate other samples (not more observations) to reduce uncertainty.



Heuristic Intuition for a Simple (Financial) Model

Consider a return stochastic model, $r_t = \mu + \sigma\varepsilon_t$, for $t = 1, 2, \cdots, T$, with (ε_t) is i.id. $\mathcal{N}(0, 1)$ [Constant Expected Return Model, CER]

$$\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} r_t \text{ and } \widehat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\big[r_t - \widehat{\mu}\big]^2$$

then (standard errors)

$$\widehat{se}[\widehat{\mu}] = \frac{\widehat{\sigma}}{\sqrt{T}} \text{ and } \widehat{se}[\widehat{\sigma}] = \frac{\widehat{\sigma}}{\sqrt{2T}}$$

then (confidence intervals)

$$\mu \in \Big[\widehat{\mu} \pm 2\widehat{se}[\widehat{\mu}]\Big] \text{ and } \sigma \in \Big[\widehat{\sigma} \pm 2\widehat{se}[\widehat{\sigma}]\Big]$$

What if the quantity of interest, θ, is another quantity, e.g. a Value-at-Risk ?

Heuristic Intuition for a Simple (Financial) Model

One can use nonparametric bootstrap

1. resampling: generate B "bootstrap samples" by resampling with replacement in the original data,

$$\boldsymbol{r}^{(b)} = \{r_1^{(b)}, \cdots, r_T^{(b)}\}, \text{ with } r_t^{(b)} \in \{r_1, \cdots, r_T\}.$$

2. For each sample $\boldsymbol{r}^{(b)}$, compute $\widehat{\theta}^{(b)}$

3. Derive the empirical distribution of $\widehat{\theta}$ from $\{\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(B)}\}$.

4. Compute any quantity of interest, standard error, quantiles, etc.

E.g. estimate the bias

$$\text{bias}[\widehat{\theta}] = \underbrace{\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}}_{\text{bootstrap mean}} - \underbrace{\frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}}_{\text{estimate}}$$

Heuristic Intuition for a Simple (Financial) Model

E.g. estimate the standard error

$$\text{se}[\widehat{\theta}] = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^{(b)} - \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^{(b)}\right)^2}$$

E.g. estimate the confidence interval, if the bootstrap distribution looks Gaussian

$$\theta \in \left[\widehat{\theta} \pm 2\text{se}[\widehat{\theta}]\right]$$

and if the distribution does not look Gaussian

$$\theta \in \left[q_{\alpha/2}^{(B)}; q_{1-\alpha/2}^{(B)}\right]$$

where $q_{\alpha}^{(B)}$ denote a quantile from $\{\widehat{\theta}^{(1)}, \cdots, \widehat{\theta}^{(B)}\}$.

Estimating the bias of $\widehat{\theta}$

Consider some statistic $\widehat{\theta}(\boldsymbol{y})$ (define on a sample $\boldsymbol{y}$).

$$\hat{\theta}_{(\cdot)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{(b)} \text{ where } \hat{\theta}_{(b)} = \hat{\theta}(\boldsymbol{y}^{(b)})$$

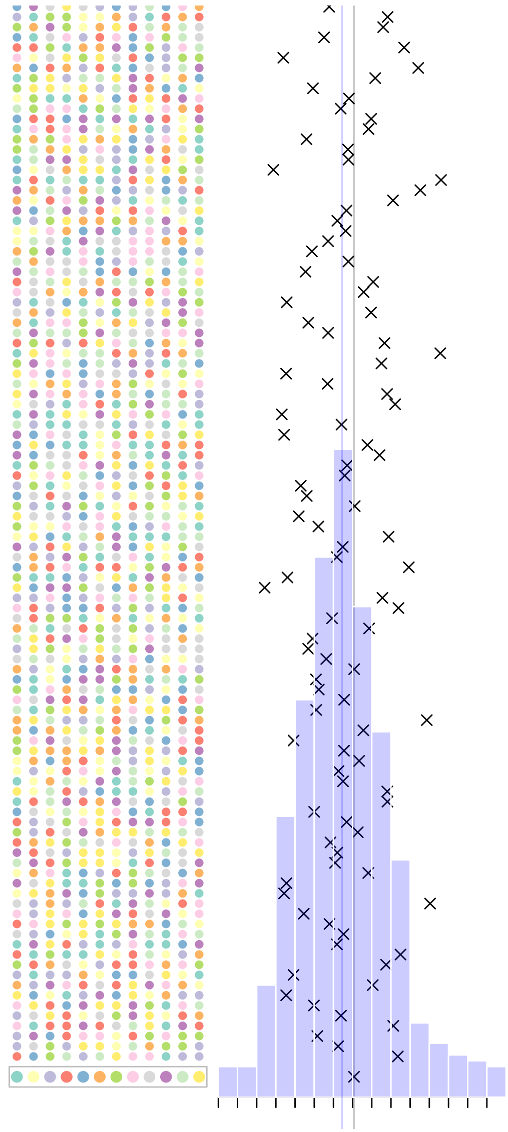
Recall that $\text{Bias}[\hat{\boldsymbol{\theta}}] = \mathbb{E}[\hat{\boldsymbol{\theta}}] - \theta$

the bootstrap estimate of the bias of the estimator $\hat{\boldsymbol{\theta}}$ is obtained by replacing $\mathbb{E}[\hat{\boldsymbol{\theta}}]$ with $\hat{\theta}_{(\cdot)}$ and θ with $\hat{\theta}$:

$$\text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\theta}_{(\cdot)} - \hat{\theta}$$

Then, since $\theta = E[\hat{\boldsymbol{\theta}}] - \text{Bias}[\hat{\boldsymbol{\theta}}]$, the bootstrap bias corrected estimate is

$$\hat{\theta}_{\text{bs}} = \hat{\theta} - \text{Bias}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \hat{\theta} - (\hat{\theta}_{(\cdot)} - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_{(\cdot)}$$

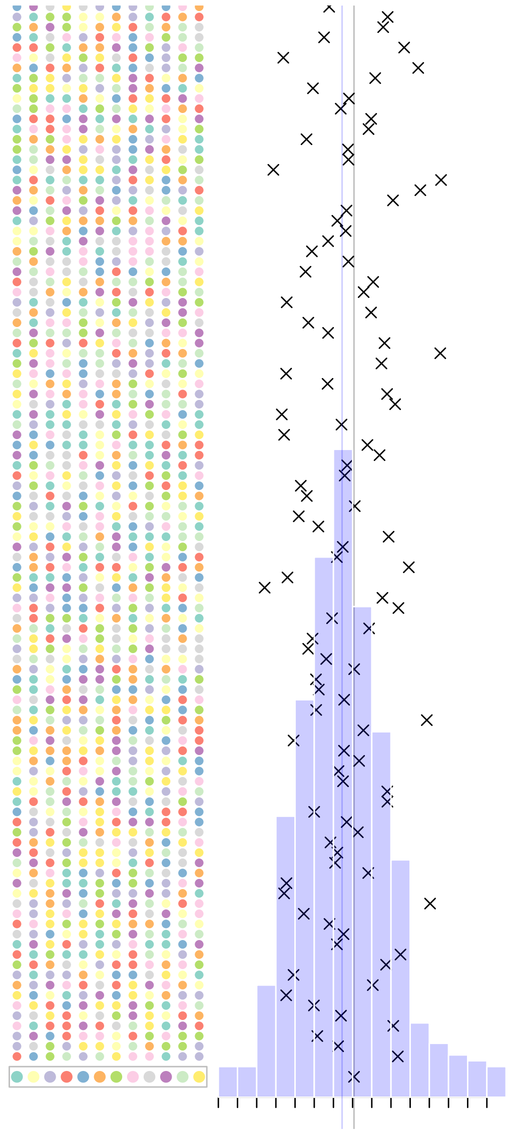


Estimating the variance of $\widehat{\theta}$

Consider some statistic $\widehat{\theta}(\boldsymbol{y})$ (define on a sample $\boldsymbol{y}$). The bootstrap approach computes the variance of the estimator $\hat{\boldsymbol{\theta}}$ through the variance of the set $\hat{\theta}_{(b)}$, $b = 1, \ldots, B$, given by

$$\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}] = \frac{\sum_{b=1}^{B}(\hat{\theta}_{(b)} - \hat{\theta}_{(\cdot)})^2}{(B-1)} \text{ where } \hat{\theta}_{(\cdot)} = \frac{\sum_{b=1}^{B}\hat{\theta}_{(b)}}{B}$$

If $\hat{\theta} = \hat{\mu}$, then for $B \to \infty$, the bootstrap estimate $\text{Var}_{\text{bs}}[\hat{\boldsymbol{\theta}}]$ converges to the variance $\text{Var}\,[\hat{\boldsymbol{\mu}}]$ (CLT).



Monte Carlo Techniques in Statistics

Law of large numbers (---), if $\mathbb{E}[X]=0$ and $\text{Var}[X]=1$: $\sqrt{n}\,\overline{X}_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

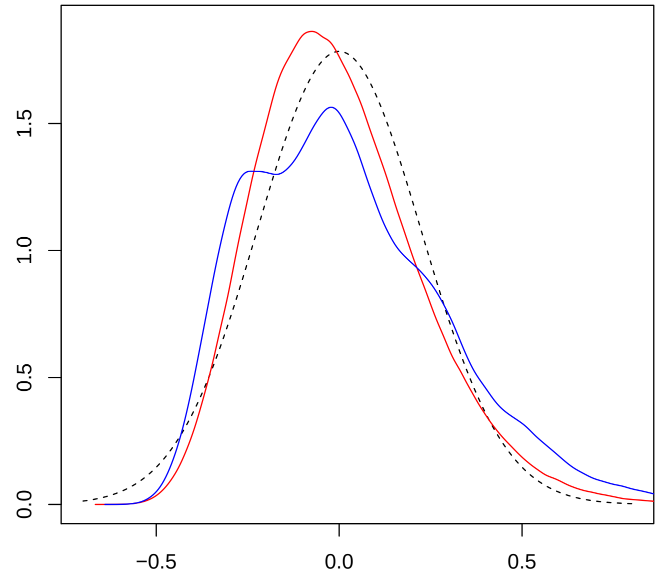
What if n is small? What is the distribution of $\overline{X}_n$?

Example : X such that $2^{-\frac{1}{2}}(X-1)\sim\chi^2(1)$

Use Monte Carlo Simulation to derive confidence interval for $\overline{X}_n$ (—).

Generate samples $\{x_1^{(m)},\cdots,x_n^{(m)}\}$ from $\chi^2(1)$, and compute $\overline{x}_n^{(m)}$

Then estimate the density of $\{\overline{x}_n^{(1)},\cdots,\overline{x}_n^{(m)}\}$, quantiles, etc.



Problem : need to know the true distribution of X.

What if we have only $\{x_1,\cdots,x_n\}$?

Generate samples $\{x_1^{(m)},\cdots,x_n^{(m)}\}$ from $\widehat{F}_n$, and compute $\overline{x}_n^{(m)}$ (—)

```
> n = 20
> ns = 1e6
> xbar = rep(NA,ns)
> for(i in 1:ns){
+    x = (rchisq(n,df=1)-1)/sqrt(2)
+    xbar[i] = mean(x)
+ }
> u = seq(-.7,.8,by=.001)
> v = dnorm(u,sd=1/sqrt(20))
> plot(u,v,col="black")
> lines(density(xbar),col="red")
> set.seed(1)
> x = (rchisq(n,df=1)-1)/sqrt(2)
> for(i in 1:ns){
+    xs = sample(x,size=n,replace=TRUE)
+    xbar[i] = mean(xs)
+ }
> lines(density(xbar),col="blue")
```

Quantifying Bias

Consider X with mean $\mu = \mathbb{E}(X)$. Let $\theta = \exp[\mu]$, then $\widehat{\theta} = \exp[\overline{x}]$ is a biased estimator of θ, see Horowitz (1998) The Bootstrap

Idea 1 : Delta Method, i.e. if $\sqrt{n}[\widehat{\tau}_n - \tau] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$, then, if $g'(\tau)$ exists and is non-null,

$$\sqrt{n}[g(\widehat{\tau}_n) - g(\tau)] \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2[g'(\tau)]^2)$$

so $\widehat{\theta}_1 = \exp[\overline{x}]$ is asymptotically unbiased.

Idea 2 : Delta Method based correction,

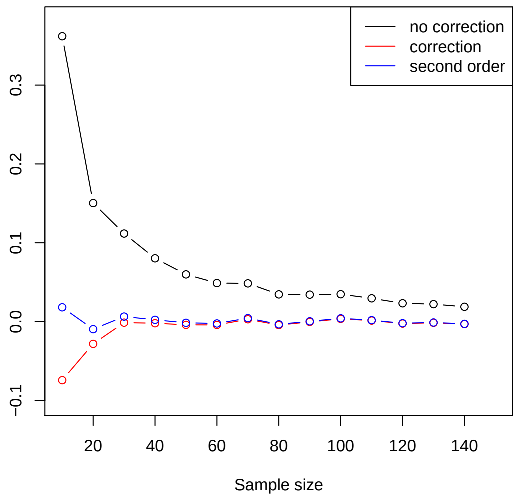
based on $\widehat{\theta}_2 = \exp\left[\overline{x} - \frac{s^2}{2n}\right]$ where $s^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}[x_i - \overline{x}]^2$.

Idea 3 : Use Bootstrap, $\widehat{\theta}_3 = \dfrac{1}{B}\displaystyle\sum_{b=1}^{n}\exp[\overline{x}^{(b)}]$

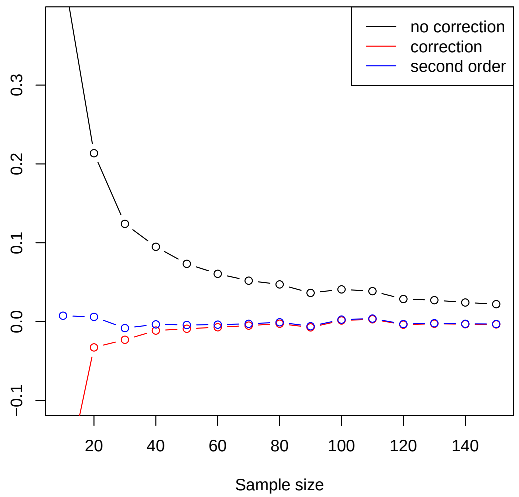
Quantifying Bias

X with mean $\mu = \mathbb{E}(X)$. Let $\theta = \exp[\mu]$. Consider three distributions

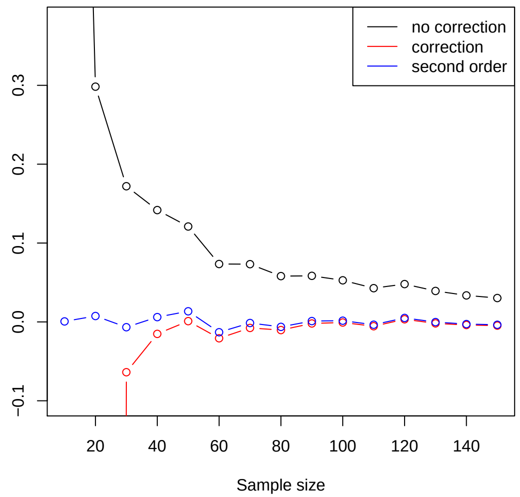
Log-normal



Student t_{10}



Student t_5

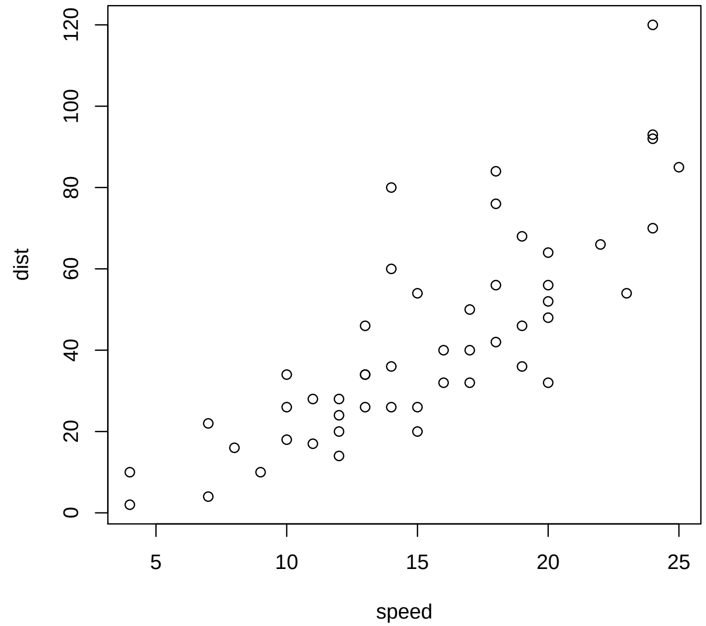


```r
> VS=matrix(NA,15,3)
> for(s in 1:15){
+ simu=function(n = 10){
+ get_i = function(i){
+    x = rnorm(n, sd=sqrt(6));
+    S = matrix(sample(x, size=n*
     10000, replace=TRUE),ncol
     =10000)
+    ThetaBoot = exp(colMeans(S))
+    Bias = mean(ThetaBoot)-exp(
     mean(x))
+    theta=exp(mean(x))/exp(.5*var
     (x)/n)
+    c(exp(mean(x)),exp(mean(x))-
     Bias, theta)
+ }
```

```r
# res = mclapply(1:2000, get_i,
    mc.cores=20)
+ res = lapply(1:10000, get_i)
+ res = do.call(rbind, res)
+ bias = colMeans(res-1)
+ return(bias)
+ }
+ VS[s,]=simu(10*s)
+ }
```

Linear Regression & Bootstrap : Parametric

1. sample $\widetilde{\varepsilon}_1^{(s)},\cdots,\widetilde{\varepsilon}_n^{(s)}$ randomly from $\mathcal{N}(0,\widehat{\sigma})$
2. set $y_i^{(s)}=\widehat{\beta}_0+\widehat{\beta}_1x_i+\widetilde{\varepsilon}_i^{(s)}$
3. consider dataset $(x_i,y_i^{(b)})=(x_i,y_i^{(b)})$'s and fit a linear regression
4. let $\widehat{\beta}_0^{(s)},\widehat{\beta}_1^{(s)}$ and $\widehat{\sigma}^{2(s)}$ denote the estimated values



Linear Regression & Bootstrap : Residuals

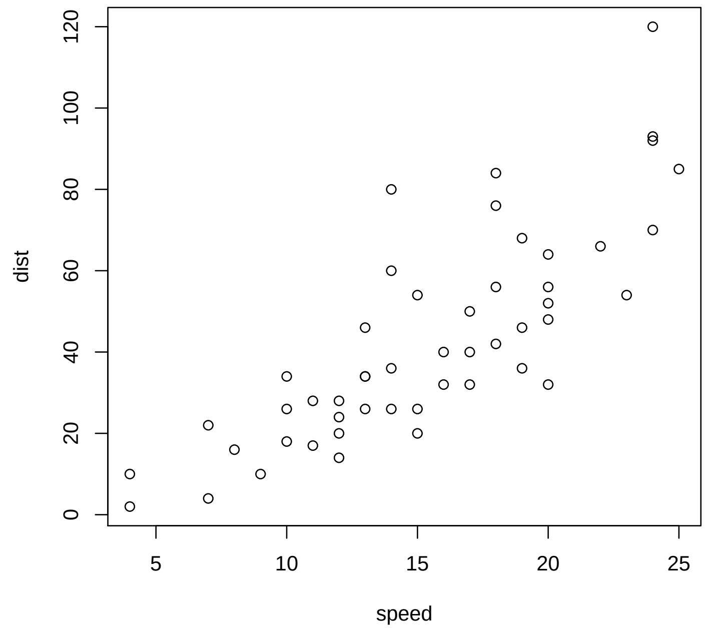
Algorithm 6.1. Davison & Hinkley (1997) Bootstrap Methods and Applications.

1. sample $\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)}$ randomly with replacement in $\{\widehat{\varepsilon}_1,\widehat{\varepsilon}_2,\cdots,\widehat{\varepsilon}_n\}$
2. set $y_i^{(b)}=\widehat{\beta}_0+\widehat{\beta}_1x_i+\widehat{\varepsilon}_i^{(b)}$
3. consider dataset $(x_i,y_i^{(b)})=(x_i,y_i^{(b)})$'s and fit a linear regression
4. let $\widehat{\beta}_0^{(b)},\widehat{\beta}_1^{(b)}$ and $\widehat{\sigma}^{2(b)}$ denote estimated values

$$\widehat{\beta}_1^{(b)}=\frac{\sum[x_i-\overline{x}]\cdot y_i^{(b)}}{\sum[x_i-\overline{x}]^2}=\widehat{\beta}_1+\frac{\sum[x_i-\overline{x}]\cdot\widehat{\varepsilon}_i^{(b)}}{\sum[x_i-\overline{x}]^2}$$

hence $\mathbb{E}[\widehat{\beta}_1^{(b)}]=\widehat{\beta}_1$, while

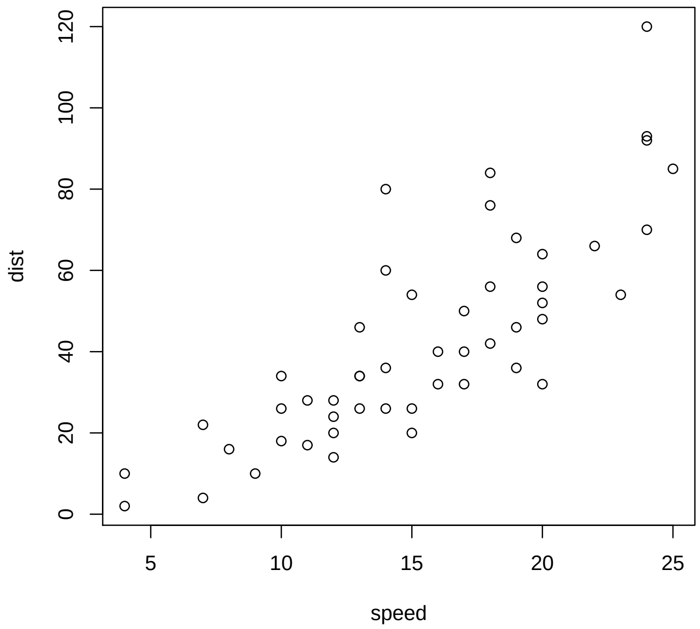
$$\text{Var}[\widehat{\beta}_1^{(b)}]=\frac{\sum[x_i-\overline{x}]^2\cdot\text{Var}[\widehat{\varepsilon}_i^{(b)}]}{\left(\sum[x_i-\overline{x}]^2\right)^2}\sim\frac{\sigma^2}{\sum[x_i-\overline{x}]^2}$$



Linear Regression & Bootstrap : Pairs

Algorithm 6.2. Davison & Hinkley (1997) Bootstrap Methods and Applications.

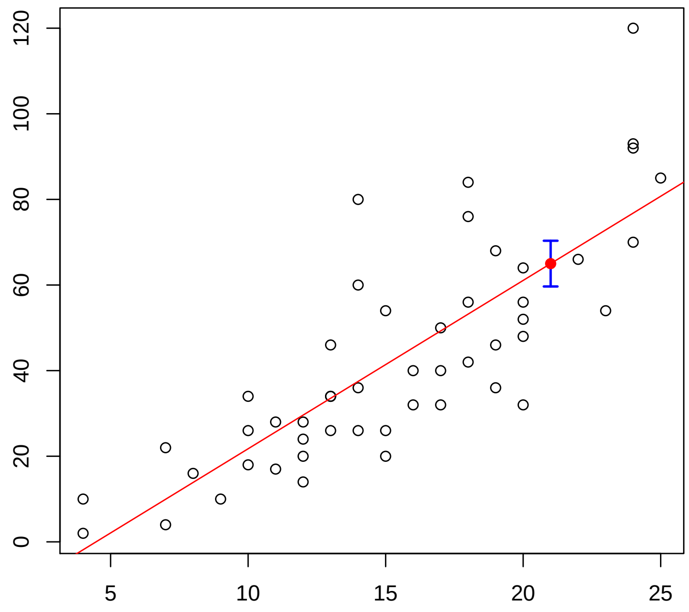
1. sample $\{i_1^{(b)},\cdots,i_n^{(b)}\}$ randomly with replacement in $\{1,2,\cdots,n\}$
2. consider dataset $(x_i^{(b)},y_i^{(b)})=(x_{i_i^{(b)}},y_{i_i^{(b)}})$'s and fit a linear regression
3. let $\widehat{\beta}_0^{(b)},\widehat{\beta}_1^{(b)}$ and $\widehat{\sigma}^{2(b)}$ denote the estimated values



Remark $\mathbb{P}(i\notin\{i_1^{(b)},\cdots,i_n^{(b)}\})=\left(1-\frac{1}{n}\right)^n\sim e^{-1}$

Key issue : residuals have to be independent and identically distributed

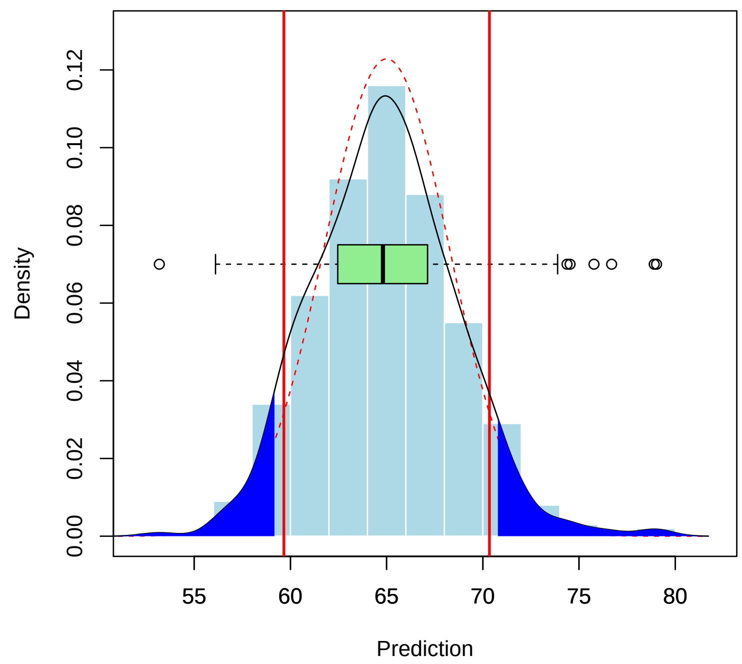
```
> plot(cars)
> reg=lm(dist~speed,data=cars)
> abline(reg,col="red")
> x=21
> predict(reg,interval="confidence",
    level=.9,newdata=data.frame(speed=x))
       fit      lwr      upr
1 65.00149 59.65934 70.34364
> Yx=rep(NA,500)
> for(s in 1:500){
+ indice=sample(1:n,size=n,replace=TRUE)
+ base=cars[indice,]
+ regb=lm(dist~speed,data=base)
+ abline(regb,col="light blue")
+ points(x,predict(regb,newdata=data.
    frame(speed=x)))
+ Yx[s]=predict(reg,newdata=data.frame(
    speed=x))
+ }
```



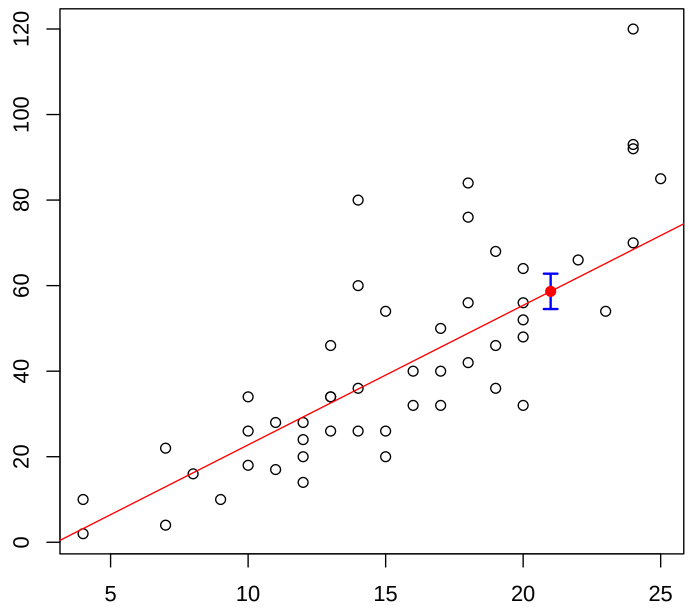
Linear Regression & Bootstrap

```

> predict(reg,interval="confidence",
    level=.9,newdata=data.frame(speed=x))
       fit      lwr      upr
1 65.00149 59.65934 70.34364
> hist(Yx,proba=TRUE)
> boxplot(Yx,horizontal=TRUE)
> lines(density(Yx))
> quantile(Yx,c(.05,.95))
      5%      95%
58.63689 70.31281
```

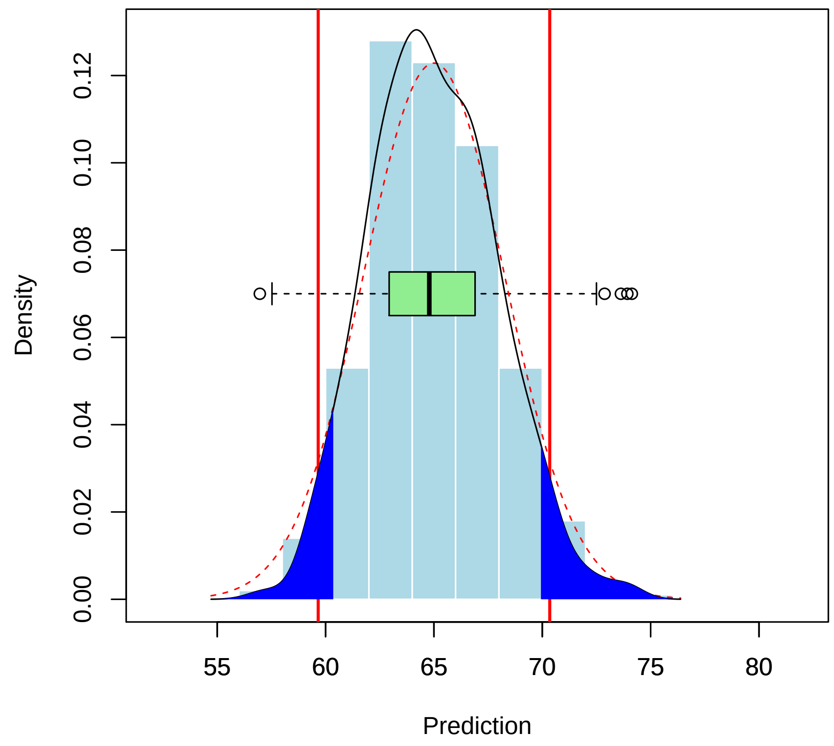


```
> plot(cars)
> reg=lm(dist~speed,data=cars)
> abline(reg,col="red")
> x=21
> predict(reg,interval="confidence",
    level=.9,newdata=data.frame(speed=x))
       fit      lwr      upr
1 65.00149 59.65934 70.34364
> base=cars
> Yx=rep(NA,500)
> for(s in 1:500){
+ indice=sample(1:n,size=n,replace=TRUE)
+ base$dist=predict(reg)+residuals(reg)[
    indice]
+ regb=lm(dist~speed,data=base)
+ abline(reg,col="light blue")
+ points(x,predict(reg,newdata=data.frame
    (speed=x)))
+ Yx[s]=predict(reg,newdata=data.frame(
    speed=x))
```



Linear Regression & Bootstrap

```
> predict(reg,interval="confidence",
     level=.9,newdata=data.frame(
   speed=x))
       fit      lwr      upr
1 65.00149 59.65934 70.34364
> hist(Yx,proba=TRUE)
> boxplot(Yx,horizontal=TRUE)
> lines(density(Yx))
```



Linear Regression & Bootstrap

Difference between the two algorithms:

1) with the second method, we make no assumption about variance homogeneity

potentially more robust to heteroscedasticity

2) the simulated samples have different designs, because the x values are randomly sampled

Key issue : residuals have to be independent and identically distributed

See discussion below on

- dynamic regression, $y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$
- heteroskedasticity, $y_i = \beta_0 + \beta_1 x_i + |x_i \cdot |\varepsilon_t$
- instrumental variables and two-stage least squares

Simulation in Econometric Models

(almost) all quantities of interest can be writen $T(\varepsilon)$ with $\varepsilon \sim F$.

E.g. $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\varepsilon}$

We need $\mathbb{E}[T(\varepsilon)] = \displaystyle\int t(\epsilon)dF(\epsilon)$

Use simulations, i.e. draw n values $\{\epsilon_1, \cdots, \epsilon_n\}$ since

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n T(\epsilon_i)\right] = \mathbb{E}[T(\varepsilon)] \text{ (unbiased)}$$

$$\frac{1}{n}\sum_{i=1}^n T(\epsilon_i) \xrightarrow{\mathcal{L}} \mathbb{E}[T(\varepsilon)] \text{ as } n \to \infty \text{ (consistent)}$$

Generating (Parametric) Distributions

Inverse cdf Technique :

Let $U \sim \mathcal{U}([0,1])$, then $X = F^{-1}(U) \sim F$.

Proof 1:

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

Proof 2: set $u = F(x)$ or $x = F^{-1}(u)$ (change of variable)

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)dF^{\star}(x) = \int_0^1 h(F^{-1}(u))du = \mathbb{E}[h(F^{-1}(U))]$$

with $U \sim \mathcal{U}([0,1])$, i.e. $X \overset{\mathcal{L}}{=} F^{-1}(U)$.

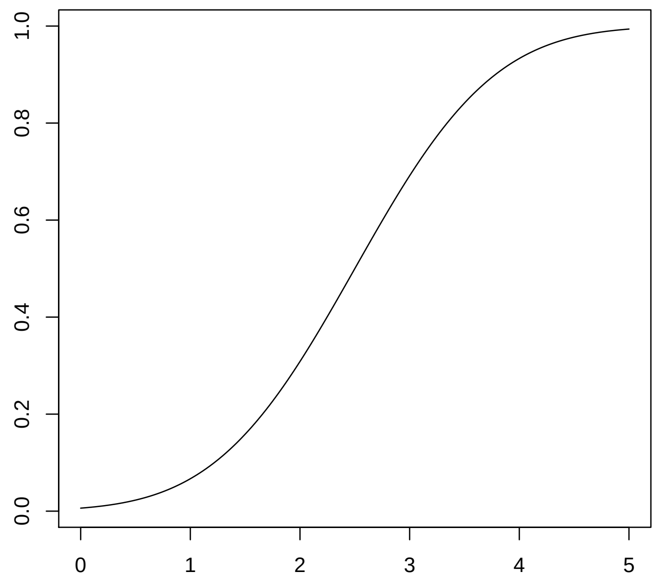
Rejection Techniques

Problem : If $X \sim F$, how to draw from $X^\star$, i.e. X conditional on $X \in [a, b]$?

Solution : draw X and use accept-reject method

1. if $x \in [a, b]$, keep it (accept)
2. if $x \notin [a, b]$, draw another value (reject)

If we generate n values, we accept - on average - $[F(b) - F(a)] \cdot n$ draws.



Importance Sampling

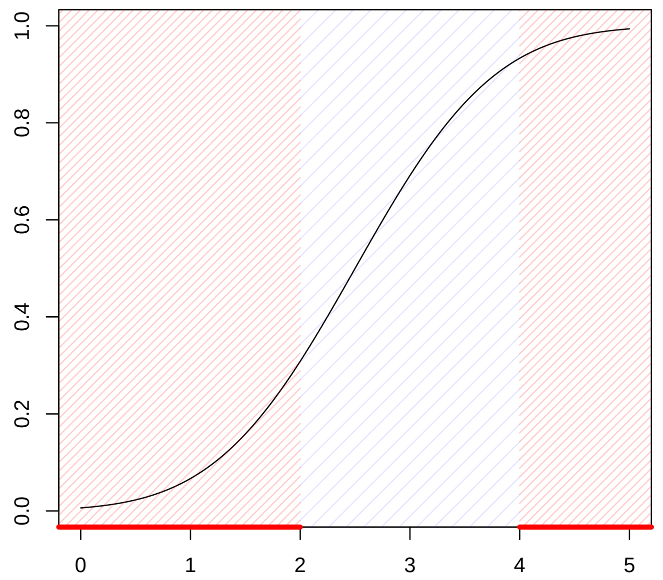
Problem : If $X \sim F$, how to draw from X conditional on $X \in [a, b]$?

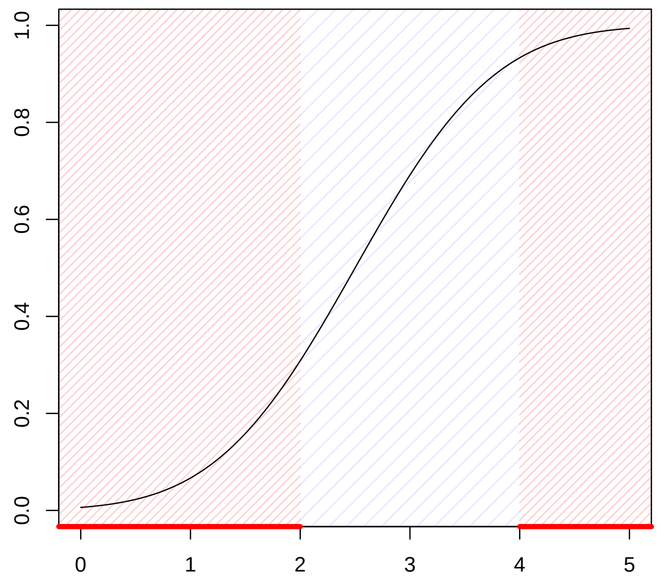
Solution : rewrite the integral and use importance sampling method

The conditional censored distribution $X^\star$ is

$$dF^\star(x) = \frac{dF(x)}{F(b) - F(a)} \mathbf{1}(x \in [a, b])$$

Alternative for truncated distributions : let $U \sim \mathcal{U}([0, 1])$ and set $\tilde{U} = [1 - U]F(a) + UF(b)$ and $Y = F^{-1}(\tilde{U})$





Going Further : MCMC

Intuition : we want to use the Central Limit Theorem, but i.id. sample is a (too) strong assumtion: if (X_i) is i.id. with distribution F,

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} h(X_i) - \int h(x)dF(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2), \text{ as } n\rightarrow\infty.$$

Use the ergodic theorem: if (X_i) is a Markov Chain with invariant measure μ,

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} h(X_i) - \int h(x)d\mu(x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2), \text{ as } n\rightarrow\infty.$$

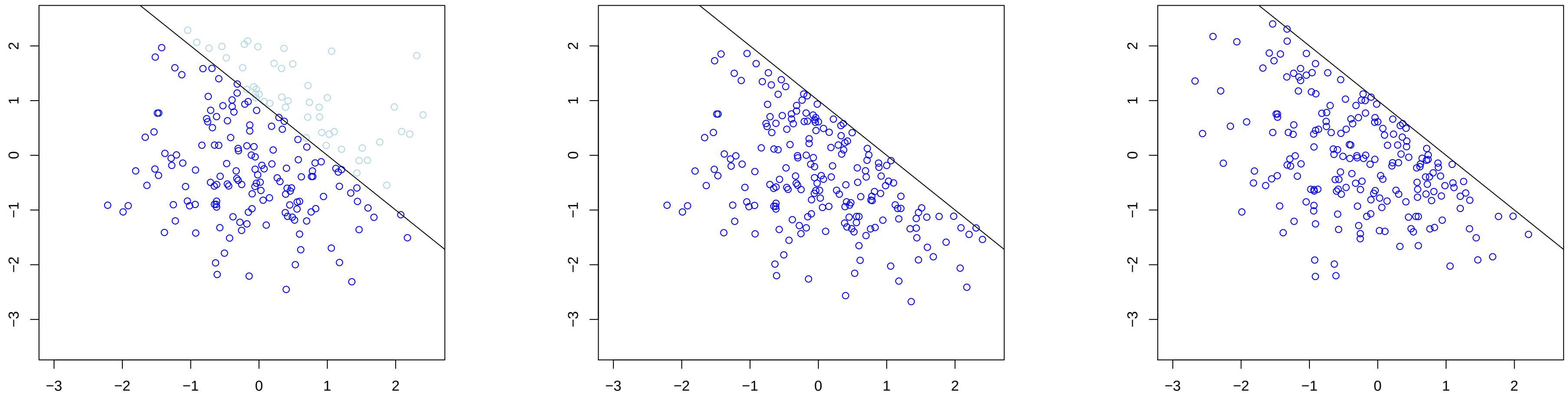
See Gibbs sampler

Example : complicated joint distribution, but simple conditional ones

Going Further : MCMC

To generate $\boldsymbol{X}|\boldsymbol{X}^{\mathsf{T}}\mathbf{1} \leq m$ with $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ (in dimension 2)

1. draw X_1 from $\mathcal{N}(0,1)$
2. draw U from $\mathcal{U}([0,1])$ and set $\tilde{U} = U\Phi(m - \epsilon_1)$
3. set $X_2 = \Phi^{-1}(\tilde{U})$



See Geweke (1991) Efficient Simulation from the Multivariate Normal and Distributions Subject to Linear Constraints

Monte Carlo Techniques in Statistics

Let $\{y_1, \cdots, y_n\}$ denote a sample from a collection of n i.id. random variables with true (unknown) distribution F_0. This distribution can be approximated by $\widehat{F}_n$.

parametric model : $F_0 \in \mathcal{F} = \{F_\theta; \theta \in \Theta\}$.

nonparametric model : $F_0 \in \mathcal{F} = \{F \text{ is a c.d.f.}\}$

The statistic of interest is $T_n = T_n(y_1, \cdots, y_n)$ (see e.g. $T_n = \widehat{\beta}_j$).

Let G_n denote the statistics of T_n:

Exact distribution : $G_n(t, F_0) = \mathbb{P}_F(T_n \leq t)$ under F_0

We want to estimate $G_n(\cdot, F_0)$ to get confidence intervals, i.e. α-quantiles

$$G_n^{-1}(\alpha, F_0) = \inf\left\{t; G_n(t, F_0) \geq \alpha\right\}$$

or p-values,

$$p = 1 - G_n(t_n, F_0)$$

Approximation of $G_n(t_n, F_0)$

Two strategies to approximate $G_n(t_n, F_0)$:

1. Use $G_\infty(\cdot, F_0)$, the asymptotic distribution as $n \to \infty$.
2. Use $G_\infty(\cdot, \widehat{F}_n)$

Here $\widehat{F}_n$ can be the empirical cdf (nonparametric bootstrap) or $F_{\widehat{\theta}}$ (parametric bootstrap).

Approximation of $G_n(t_n, F_0)$: Linear Model

Consider the test of $H_0 : \beta_j = 0$, p-value being $p = 1 - G_n(t_n, F_0)$

- Linear Model with Normal Errors $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

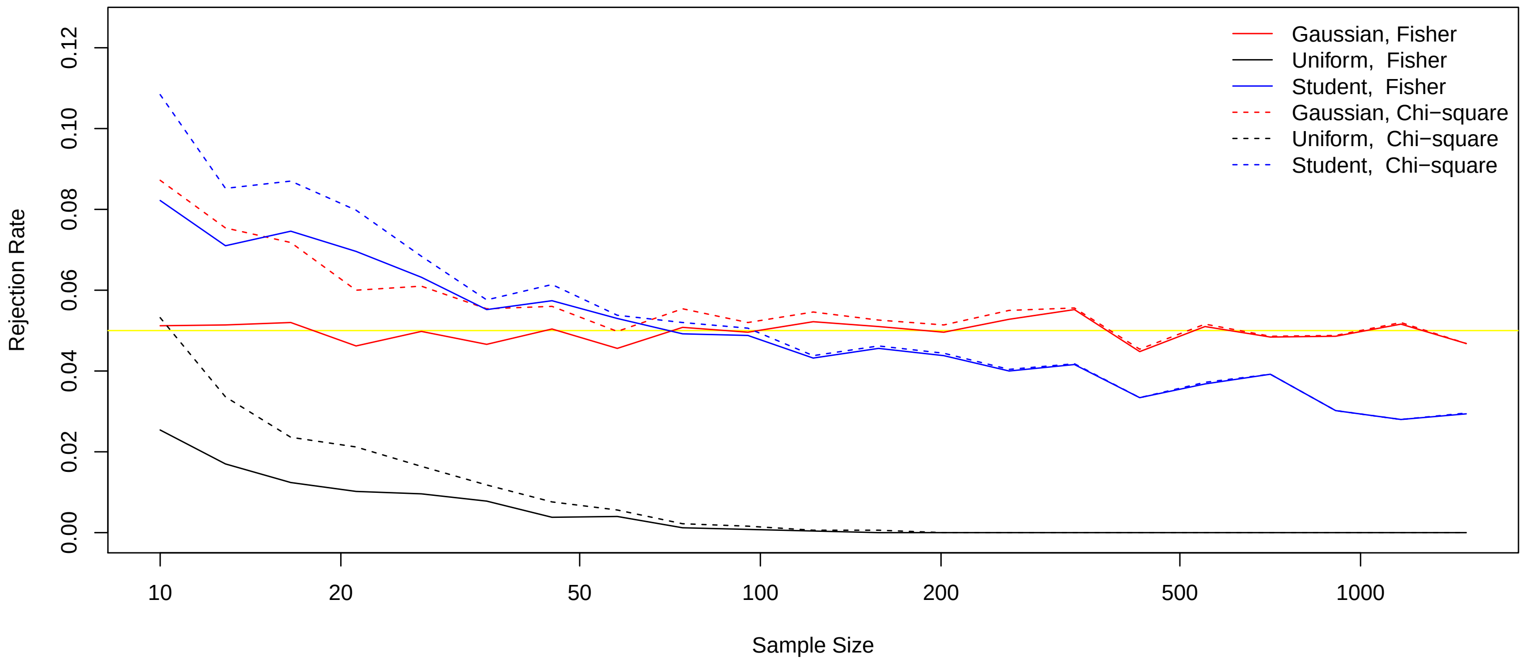
Then $\dfrac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2} \sim \mathcal{F}(1, n-k) = G_n(\cdot, F_0)$ where F_0 is $\mathcal{N}(0, \sigma^2)$

- Linear Model with Non-Normal Errors $y_i = \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta} + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i] = 0$.

Then $\dfrac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2} \xrightarrow{\mathcal{L}} \xi^2(1) = G_\infty(\cdot, F_0)$ as $n \to \infty$.

Approximation of $G_n(t_n, F_0)$: Linear Model

Application $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.



Here F_0 is $\mathcal{N}(0, \sigma^2)$

```
> pvf = function(t) mean((1-pf(t
    ,1,length(t)-2))<.05)
> pvq = function(t) mean((1-
    pchisq(t,1)<.05)
> TABLE= function(n=30){
+ ns = 5000
+ x = c(1.0001,rep(1,n-1))
+ e = matrix(rnorm(n*ns),n)
+ e2 = matrix(runif(n*ns,-3,3),n)
+ e3 = matrix(rt(n*ns,2),n)
+ get_i = function(i){
+ r1 = lm(e[,i]~x)
+ r2 = lm(e2[,i]~x)
+ r3 = lm(e3[,i]~x)
+ t1 = r1$coef[2]^2/vcov(r1)[2,2]
+ t2 = r2$coef[2]^2/vcov(r2)[2,2]
+ t3 = r3$coef[2]^2/vcov(r3)[2,2]
+ c(t1,t2,t3)}
```

```
+ library(parallel)
# t = mclapply(1:ns, get_i, mc.
    cores=50)
+ t = lapply(1:ns, get_i)
+ t = simplify2array(t)
+ rj1 = pvf(t[,1])
+ rj2 = pvf(t[,2])
+ rj3 = pvf(t[,3])
+ rj12 = pvq(t[,1])
+ rj22 = pvq(t[,2])
+ rj32 = pvq(t[,3])
+ ans = rbind(c(rj1,rj2,rj3),c(
    rj12,rj22,rj32))
+ return(ans) }
> TABLE(30)
```

Approximation of $G_n(t_n, F_0)$: Linear Model

```
> ns=1e5
> PROP=matrix(NA,ns,6)
> n=30
> VN=seq(10,140,by=10)
> for(s in 1:ns){
+ X=rnorm(n)
+ E=rnorm(n)
+ Y=1+X+E
+ reg=lm(Y~X)
+ T=(coefficients(reg)[2]-1)^2/
    vcov(reg)[2,2]
+ PROP[s,1]=T>qf(.95,1,n-2)
+ PROP[s,2]=T>qchisq(.95,1)
+ E=rt(n,df=3)
+ Y=1+X+E
```

```
+ reg=lm(Y~X)
+ T=(coefficients(reg)[2]-1)^2/
    vcov(reg)[2,2]
+ PROP[s,3]=T>qf(.95,1,n-2)
+ PROP[s,4]=T>qchisq(.95,1)
+ E=runif(n)*4-2
+ Y=1+X+E
+ reg=lm(Y~X)
+ T=(coefficients(reg)[2]-1)^2/
    vcov(reg)[2,2]
+ PROP[s,5]=T>qf(.95,1,n-2)
+ PROP[s,6]=T>qchisq(.95,1)
+ }
> apply(PROP,mean,2)
```

Computation of $G_\infty(t,\widehat{F}_n)$

For $b\in\{1,\cdots,B\}$, generate boostrap samples of size n, $\{\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)}\}$ by drawing from $\widehat{F}_n$.

Compute $T^{(b)}=T_n(\widehat{\varepsilon}_1^{(b)},\cdots,\widehat{\varepsilon}_n^{(b)})$, and use sample $\{T^{(1)},\cdots,T^{(B)}\}$ to compute $\widehat{G}$,

$$\widehat{G}(t)=\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(T^{(b)}\leq t)$$

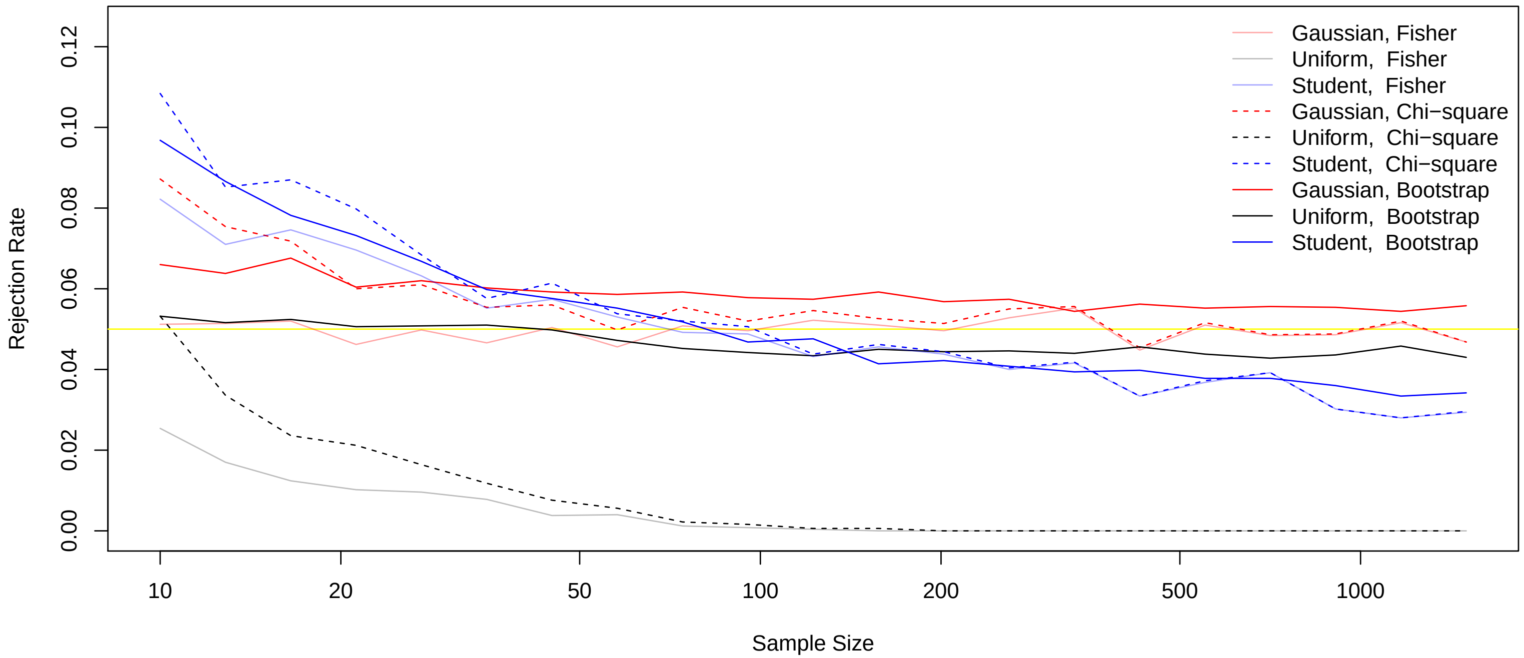
Linear Model: computation of $G_\infty(t,\widehat{F}_n)$

Consider the test of $H_0 : \beta_j = 0$, p-value being $p = 1 - G_n(t_n, F_0)$

1. compute $t_n = \dfrac{(\widehat{\beta}_j - \beta_j)^2}{\widehat{\sigma}_j^2}$
2. generate B boostrap samples, under the null assumption
3. for each boostrap sample, compute $t_n^{(b)} = \dfrac{(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j)^2}{\widehat{\sigma}_j^{2(b)}}$
4. reject H_0 if $\dfrac{1}{B}\displaystyle\sum_{i=1}^{B} \mathbf{1}(t_n > t_n^{(b)}) < \alpha$.

Linear Model: computation of $G_\infty(t, \widehat{F}_n)$

Application $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.



```
> TABLE2=function(n=30){
+ B = 299
+ sn = sqrt(n/(n-1))
+ ns = 5000
+ x = rep(1,n)
+ x[1] = 1.0001
+ e = matrix(rnorm(n*ns),n)
+ e2 = matrix(runif(n*ns,-3,3),n)
+ e3 = matrix(rt(n*ns,2),n)
+ b0tilde1 = colMeans(e)
+ b0tilde2 = colMeans(e2)
+ b0tilde3 = colMeans(e3)
+ getB_i = function(i){
+ u1 = (e[,i]-b0tilde1[i])*sn
+ u2 = (e2[,i]-b0tilde2[i])*sn
+ u3 = (e3[,i]-b0tilde3[i])*sn
+ Indic = matrix(sample(n, size=B
    *n, replace=TRUE), n, B)
```

```
+ y1 = u1[Indic[,j]]+b0tilde1[i]

+ y2 = u2[Indic[,j]]+b0tilde2[i]
+ y3 = u3[Indic[,j]]+b0tilde3[i]

+ r1 = lm(y1~x)
+ r2 = lm(y2~x)
+ r3 = lm(y3~x)
+ t = r1$coef[2]^2/vcov(r1)[2,2]
+ t2 = r2$coef[2]^2/vcov(r2)[2,2]
+ t3 = r3$coef[2]^2/vcov(r3)[2,2]
+ c(t,t2,t3) }
+  res = sapply(1:B, getB_j)
+  rj1 = mean(res[1,]<t[1,i])
+  rj2 = mean(res[2,]<t[2,i])
+  rj3 = mean(res[3,]<t[3,i])
+  c(rj1, rj2, rj3)<0.05 }
+  tstar = lapply(1:ns, getB_i)
+  tstar = simplify2array(tstar)
+  tstar <- rowMeans(tstar)
```

Linear Regression

What does generate B boostrap samples, under the null assumption means ?

Use residual boostrap technique:

Example : (standard) linear model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under H_0, i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta} = \sqrt{\dfrac{n}{n-1}}\widehat{\eta}$

2.3. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

2.4. Set $y_i^{(b)} = \widehat{\beta}_0 + \widetilde{\eta}_i^{(b)}$

2.5. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

Going Further on Linear Regression

Recall that the OLS estimator satisfies

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i$$

while for the bootstrap

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i^{(b)}$$

Thus, for i.id. data, the variance is

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{X}_i\varepsilon_i\right)^\top\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^\top\varepsilon_i^2\right]$$

Going Further on Linear Regression

and similarly (for i.id. data)

$$\mathbb{E}\left[\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i^{(b)}\right)\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\varepsilon_i^{(b)}\right)^{\mathsf{T}}\middle|\boldsymbol{X},Y\right]=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\widehat{\varepsilon}_i^2$$

Bootstrap with dynamic regression models

Example : linear model, $y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under H_0, i.e. $y_t = \beta_0 + \beta_2 y_{t-1} + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$ (estimated residuals from an AR(1))

2.2. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta} = \sqrt{\dfrac{n}{n-2}}\widehat{\eta}$

2.3. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

2.4. Set (recursively) $y_t^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_2 y_{t-1}^{(b)} + \widetilde{\eta}_t^{(b)}$

2.5. Estimate the regression model $y_t^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_t + \beta_2^{(b)} y_{t-1}^{(b)} + \varepsilon_t^{(b)}$

Remark : start (usually) with $y_0^{(b)} = y_1$

Bootstrap with heteroskedasticity

Example : linear model, $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_t$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under H_0, i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Compute $H_{i,i}$ with $\boldsymbol{H} = [H_{i,i}]$ from $\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{X}^\top \boldsymbol{X}]^{-1}\boldsymbol{X}^\top$.

2.3.a. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta}_i = \pm \dfrac{\widehat{\eta}_i}{\sqrt{1 - H_{i,i}}}$

(here $\pm$ mean $\{-1, +1\}$ with probabilities $\{1/2, 1/2\}$)

2.4.a. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

2.5.a. Set $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_2 y_{i-1}^{(b)} + \widetilde{\eta}_i^{(b)}$

2.6.a. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

This was suggested in Liu (1988) Bootstrap procedures under some non - i.i.d. models

Bootstrap with heteroskedasticity

Example : linear model, $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_t$ with $H_0 : \beta_1 = 0$.

2.1. Estimate the model under H_0, i.e. $y_i = \beta_0 + \eta_i$, and save $\{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Compute $H_{i,i}$ with $\boldsymbol{H} = [H_{i,i}]$ from $\boldsymbol{H} = \boldsymbol{X}[\boldsymbol{X}^\top \boldsymbol{X}]^{-1}\boldsymbol{X}^\top$.

2.3.b. Define $\widetilde{\boldsymbol{\eta}} = \{\widetilde{\eta}_1, \cdots, \widetilde{\eta}_n\}$ with $\widetilde{\eta}_i = \xi_i \dfrac{\widehat{\eta}_i}{\sqrt{1 - H_{i,i}}}$

(here ξ_i takes values $\left\{\dfrac{1-\sqrt{5}}{2}, \dfrac{1+\sqrt{5}}{2}\right\}$ with probabilities $\left\{\dfrac{\sqrt{5}+1}{2\sqrt{5}}, \dfrac{\sqrt{5}-1}{2\sqrt{5}}\right\}$)

2.4.b. Draw (with replacement) residuals $\widetilde{\boldsymbol{\eta}}^{(b)} = \{\widetilde{\eta}_1^{(b)}, \cdots, \widetilde{\eta}_n^{(b)}\}$

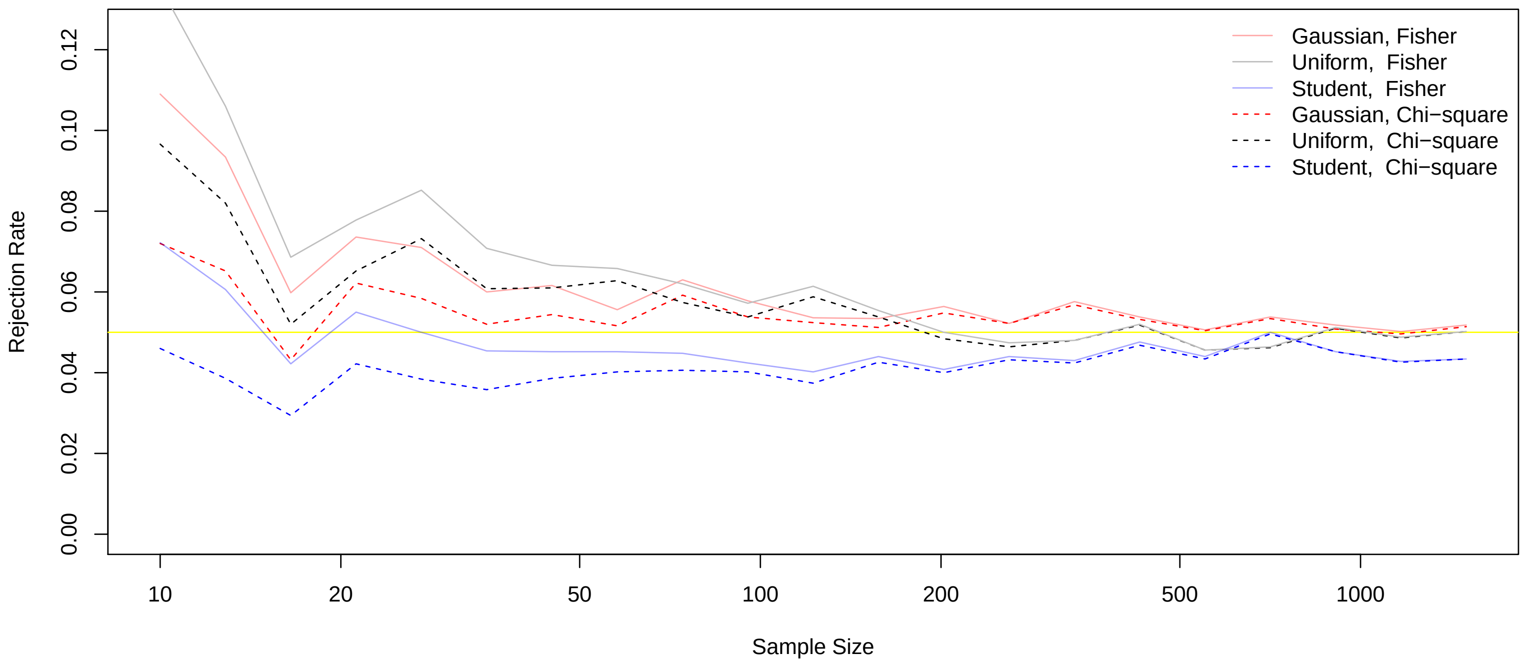
2.5.b. Set $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\beta}_2 y_{i-1}^{(b)} + \widetilde{\eta}_i^{(b)}$

2.6.b. Estimate the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i + \varepsilon_i^{(b)}$

This was suggested in Mammen (1993) Bootstrap and wild bootstrap for high dimensional linear models, ξ_i's satisfy here $\mathbb{E}[\xi_i^3] = 1$

Bootstrap with heteroskedasticity

Application $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.



Bootstrap with 2SLS: Wild Bootstrap

Consider a linear model, $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{x}_i = \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \boldsymbol{u}_i$.

Two-stage least squares:

1. regress each column of $\boldsymbol{x}$ on $\boldsymbol{z}$, $\widehat{\boldsymbol{\gamma}} = [\boldsymbol{Z}^\top \boldsymbol{Z}]\boldsymbol{Z}^\top \boldsymbol{X}$ and consider the predicted value

$$\widehat{\boldsymbol{X}} = \boldsymbol{Z}\widehat{\boldsymbol{\gamma}} = \underbrace{\boldsymbol{Z}[\boldsymbol{Z}^\top \boldsymbol{Z}]\boldsymbol{Z}^\top}_{\Pi_{\boldsymbol{Z}}} \boldsymbol{X}$$

2. regress y on predicted covariates $\widehat{\boldsymbol{X}}$, $y_i = \widehat{\boldsymbol{x}}_i^\top \boldsymbol{\beta} + \varepsilon_i$

Bootstrap with 2SLS: Wild Bootstrap

Example : linear model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_t$ where $x_i = \boldsymbol{z}_i^\mathsf{T}\boldsymbol{\gamma} + u_i$ and $\text{Cov}[\varepsilon, u] = \rho$, with $H_0 : \beta_1 = 0$.

So called Wild Boostrap, see Davidson & Mackinnon (2009) Wild bootstrap tests for IV regression

2.1. Estimate the model under H_0, i.e. $y_i = \beta_0 + \eta_i$, by 2SLS and save $\widehat{\boldsymbol{u}} = \{\widehat{\eta}_1, \cdots, \widehat{\eta}_n\}$

2.2. Estimate $\boldsymbol{\gamma}$ from $x_i = \boldsymbol{z}_i^\mathsf{T}\boldsymbol{\gamma} + \delta\widehat{\eta}_i + u_i$

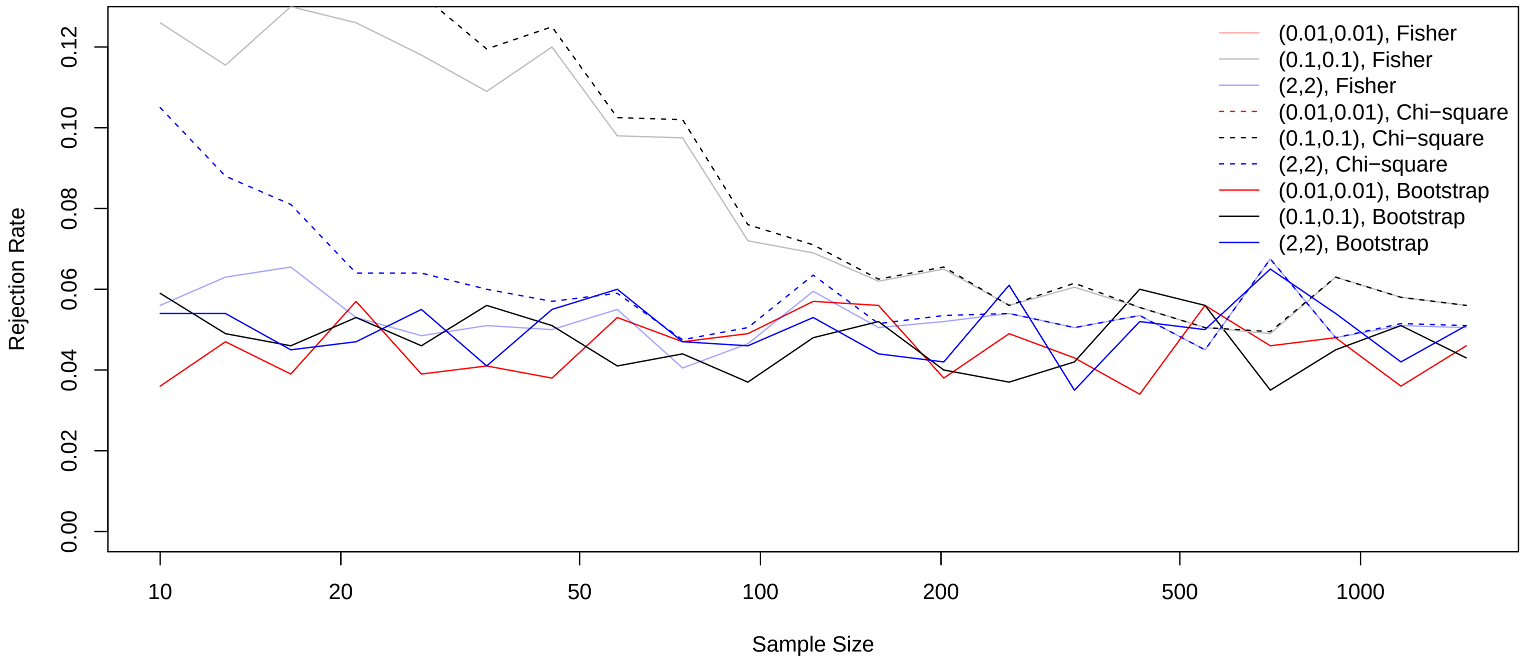
2.3. Define $\widetilde{\boldsymbol{u}} = \{\widetilde{u}_1, \cdots, \widetilde{u}_n\}$ with $\widetilde{u}_i = X_i - \boldsymbol{z}_i^\mathsf{T}\widehat{\boldsymbol{\gamma}}$

2.4. Draw (with replacement) pairs of residuals $(\widehat{\boldsymbol{\eta}}^{(b)}, \widetilde{\boldsymbol{u}}^{(b)})$ of $(\widehat{\eta}_i^{(b)}, \widetilde{u}_i^{(b)})$'s

2.5. Set $x_i^{(b)} = \boldsymbol{z}_i^\mathsf{T}\widehat{\boldsymbol{\gamma}} + \widetilde{u}_i^{(b)}$ and $y_i^{(b)} = \widehat{\beta}_0 + \widehat{\eta}_i^{(b)}$

2.6. Estimate (using 2SLS) the regression model $y_i^{(b)} = \beta_0^{(b)} + \beta_1^{(b)} x_i^{(b)} + \varepsilon_i^{(b)}$, where $x_i^{(b)} = \boldsymbol{z}_i^\mathsf{T}\boldsymbol{\gamma} + u_i$

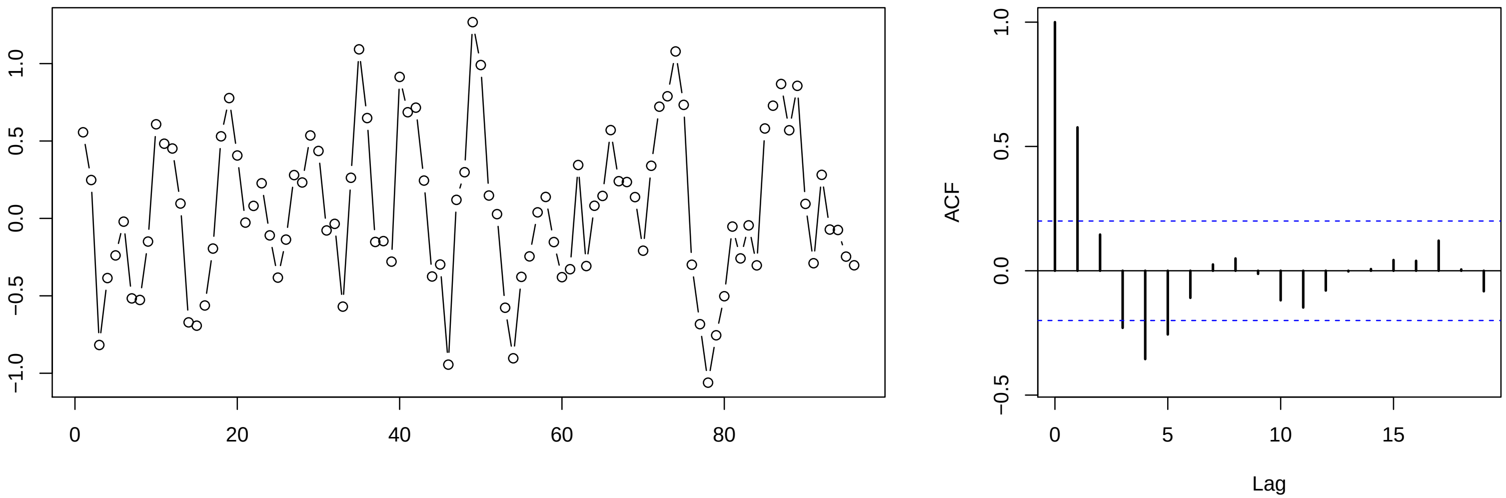
Bootstrap with 2SLS: Wild Bootstrap



See example Section 5.2 in Horowitz (1998) The Bootstrap.

Bootstrap for time series

Application $y_i = \beta_0 + \beta_1 x_i + |x_i| \cdot \varepsilon_i$, $\varepsilon \sim \mathcal{N}(0,1)$, $\varepsilon \sim \mathcal{U}([-1,+1])$ or $\varepsilon \sim \mathcal{S}td(\nu = 2)$.



The strategy is to estimate the predictive model, to derive the (i.i.d.) residuals, and to bootstrap the residuals.

Bootstrap for time series

Consider a linear $AR(p)$ (causal) time series model,

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$

such that (ε_t) is a white noise. Fit the model and derive estimated residuals,

$$y_t = \widehat{\varphi}_1 y_{t-1} + \widehat{\varphi}_2 y_{t-2} + \cdots + \widehat{\varphi}_p y_{t-p} + \widehat{\varepsilon}_t$$

Draw a bootstraped sample $\{\widehat{\varepsilon}_1^{(b)}, \cdots, \widehat{\varepsilon}_T^{(b)}\}$, and iteratively

$$y_t^{(b)} = \widehat{\varphi}_1 y_{t-1}^{(b)} + \widehat{\varphi}_2 y_{t-2}^{(b)} + \cdots + \widehat{\varphi}_p y_{t-p}^{(b)} + \widehat{\varepsilon}_t^{(b)}$$

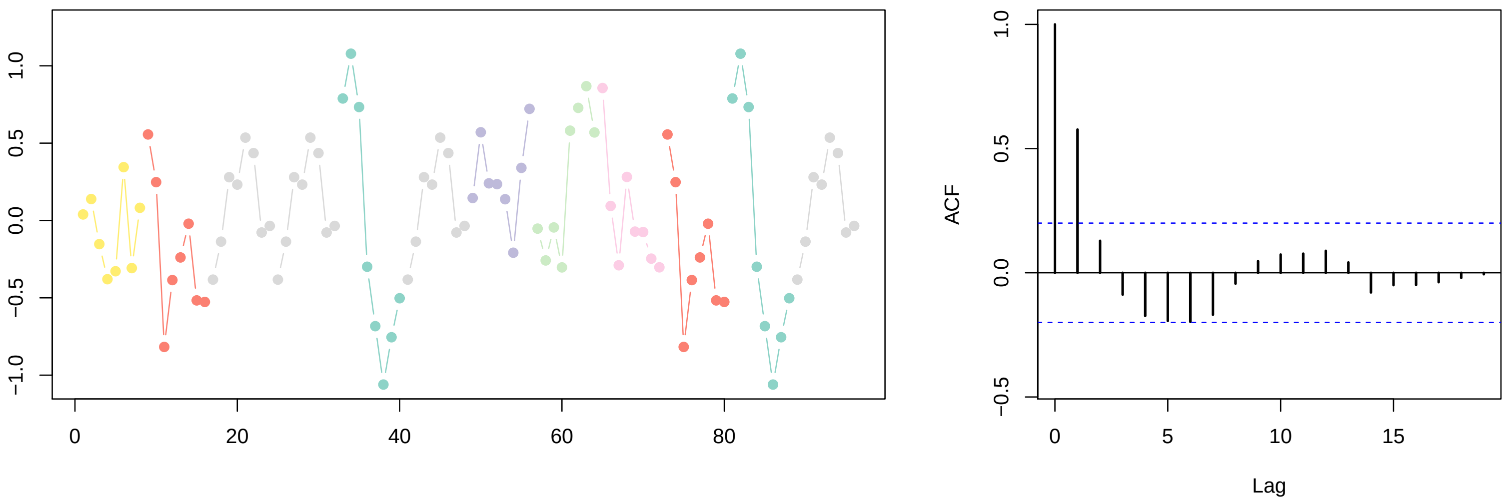
Let $B_{t,\ell} = \{y_t, \cdots, y_{t+l}\}$ denote the block starting in t of length ℓ.

Consider a (very) simple time series, $\{y_1, y_2, y_3, y_4\}$ to illustrate.

Bootstrap for time series

non-overlapping block bootstrap, from Carlstein (1986). Assume that $T = k\ell$, so that there are $k = T/\ell$ (non-overlapping) blocks of length ℓ. For instance, with blocks of length 2, blocks are $\{(y_1, y_2), (y_3, y_4)\}$. Let $I_1, \cdots, I_k$ denote random variables uniformly distributed on $\{1, \ell+1, \cdots, (k-1)\ell+1\}$, then

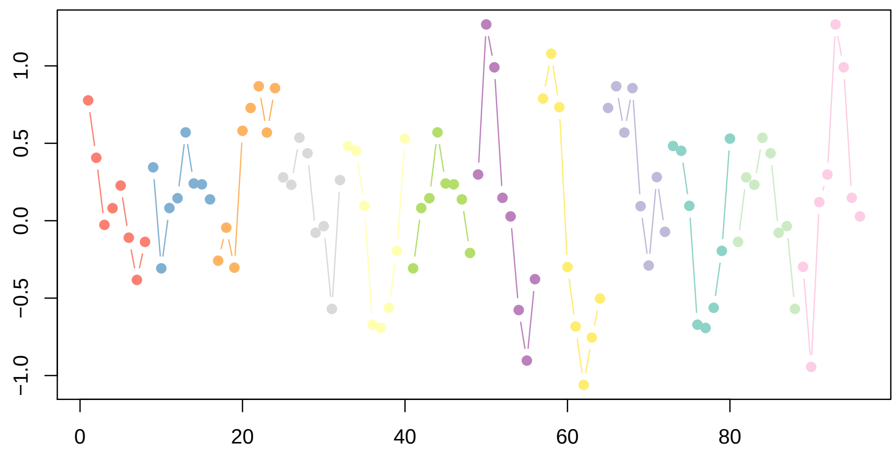
$$\boldsymbol{y}^{(b)} = \{B(I_1,\ell), B(I_2,\ell), \cdots, B(I_{T/\ell},\ell)\}$$

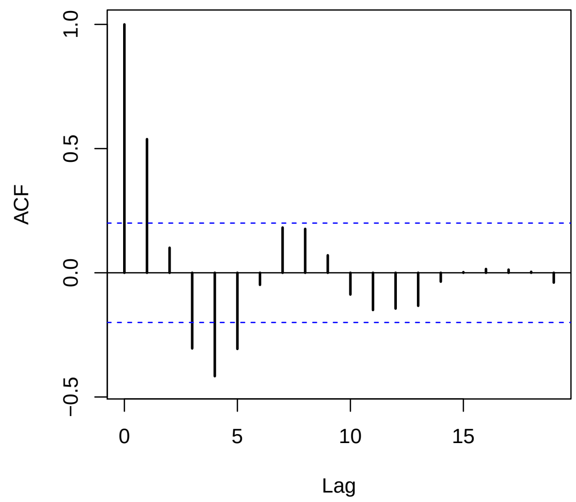


Bootstrap for time series

moving block bootstrap, from Kunsch (1989) or Liu & Singh (1992). For instance, with blocks of length 2, blocks are $\{(y_1, y_2), (y_2, y_3), (y_3, y_4)\}$. There are $k = T - \ell + 1$ blocks here. Let $I_1, \cdots, I_{T/\ell}$ denote random variables uniformly distributed on $\{1, 2, \cdots, (k-1)\ell + 1\}$, then

$$\boldsymbol{y}^{(b)} = \{B(I_1, \ell), B(I_2, \ell), \cdots, B(I_{T/\ell}, \ell)\}$$





Bootstrap for time series

circular block bootstrap, from Politis & Romano (1992). Assume the data live on a circle so that $y_{T+1}=y_1$, etc. For instance, with blocks of length 2, blocks are

$$\{(y_1,y_2),(y_2,y_3),(y_3,y_4),(y_4,y_1)\}$$

Let $I_1,\cdots,I_{T/\ell}$ denote random variables uniformly distributed on $\{1,2,\cdots,T\}$, then

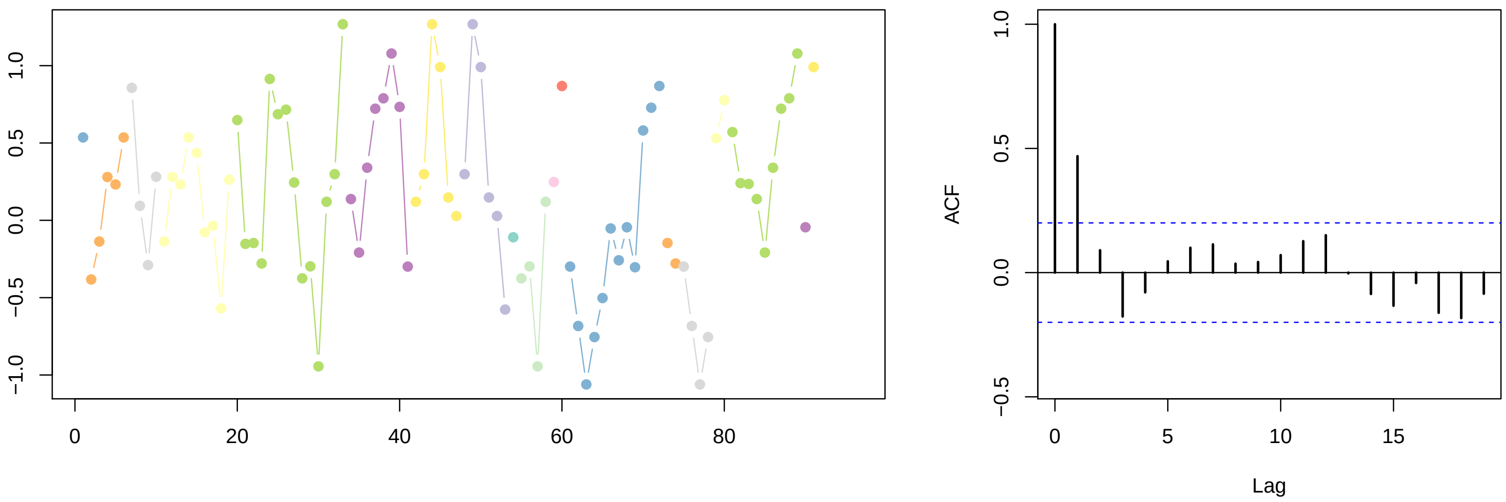
$$\boldsymbol{y}^{(b)}=\big\{B(I_1,\ell),B(I_2,\ell),\cdots,B(I_{T/\ell},\ell)\big\}$$

stationary block bootstrap, from Politis & Romano (1994), blocks have random sizes.

Bootstrap for time series

Let $L_1, \cdots, L_k$ denote geometric distributions with probability $p = \ell^{-1}$, so that $L_1 + \cdots + L_{k-1} < n$ and $L_1 + \cdots + L_{k-1} + L_k \geq n$. Let $I_1, \cdots, I_k$ denote random variables uniformly distributed on $\{1, 2, \cdots, n\}$, then

$$\boldsymbol{y}^{(b)} = \{B(I_1, L_1), B(I_2, L_2), \cdots, B(I_k, L_k)\}$$



Bootstrap for panel data

Consider some panel dataset, say $(y_{i,t})$, with n individuals (i) observed over T period of time (t). Let $\boldsymbol{Y}$ denote the matrix of observations.

Several models can be considered

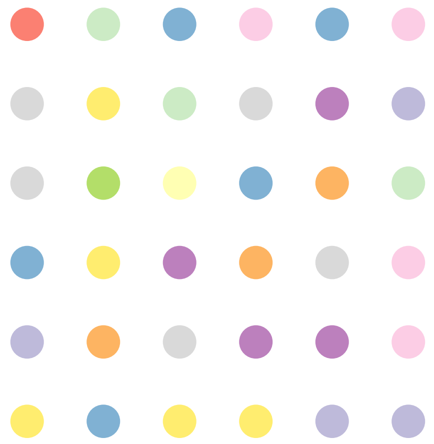
$$y_{i,t} = \theta + \varepsilon_{i,t} \tag{i.i.d.}$$

$$y_{i,t} = \theta + \alpha_i + \varepsilon_{i,t} \text{ or } y_{i,t} = \theta + \beta_t + \varepsilon_{i,t} \tag{one-way}$$

$$y_{i,t} = \theta + \alpha_i + \beta_t + \varepsilon_{i,t} \tag{two-way}$$

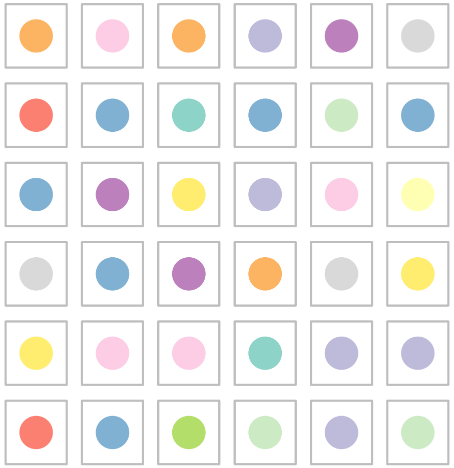
$$y_{i,t} = \theta + \alpha_i F_t \varepsilon_{i,t} \tag{factor}$$

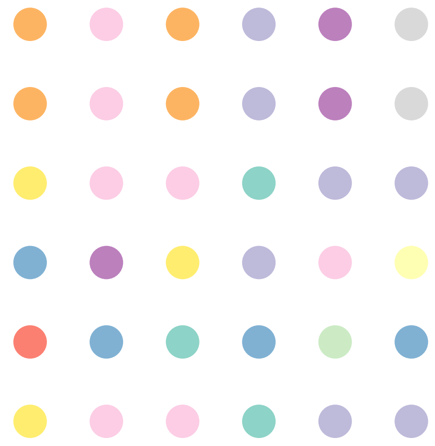
for some (common) factor (F_t).



Bootstrap for panel data

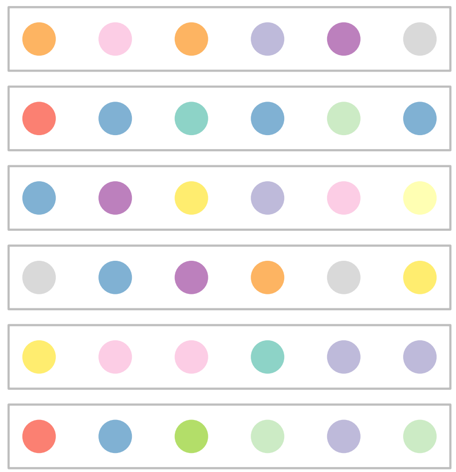
The bootrap technique described previously is refered to i.i.d. bootstrap: each observation $y_{i,t}$ is drawn (with replacement) from values in $\boldsymbol{Y}$ (with equal probabilty $1/nT$).

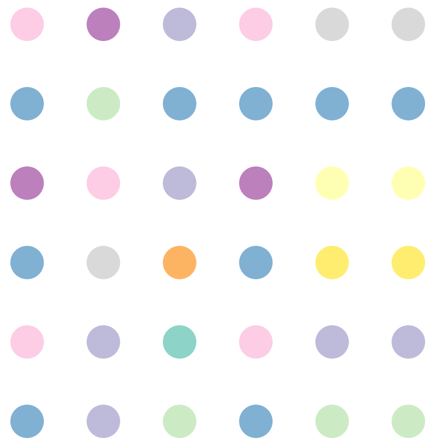




Bootstrap for panel data

One can consider individual bootstrap by sampling rows $\boldsymbol{y}_i = (y_{i,1}, \cdots, y_{i,T})$ from rows of $\boldsymbol{Y}$ (with equal probabilty $1/n$).

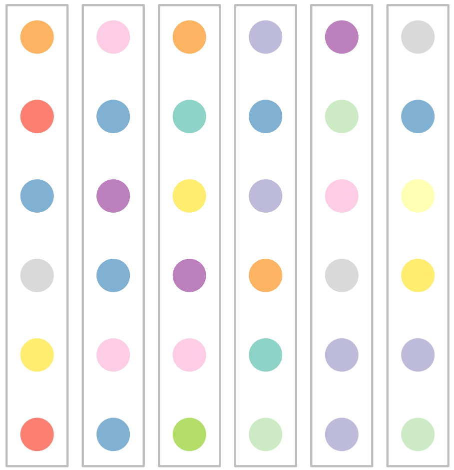




Bootstrap for panel data

One can consider temporal bootstrap by sampling columns $\boldsymbol{y}_t = (y_{1,t}, \cdots, y_{n,t})$ from columns of $\boldsymbol{Y}$ (with equal probabilty $1/T$).

One can consider block bootstrap. Assume that $T = k\ell$, so that there are k (non-overlapping) blocks of length ℓ.



Bootstrap for panel data

One can consider double bootstrap, with a combination of individual and temporal bootsrap (or block boostrap for the later).

Estimation of Various Quantities of Interest

Consider a quadratic model,

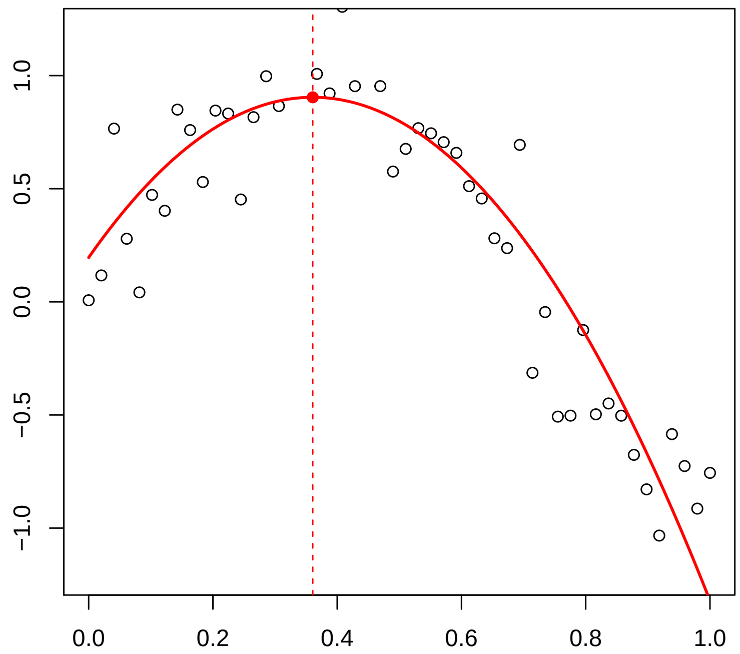
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

The minimum is obtained in $\theta = -\beta_1/2\beta_2$.

What could be the standard error for θ ?

1. Use of the Delta-Method

$\theta = g(\beta_1, \beta_2) = \dfrac{-\beta_1}{2\beta_2}$



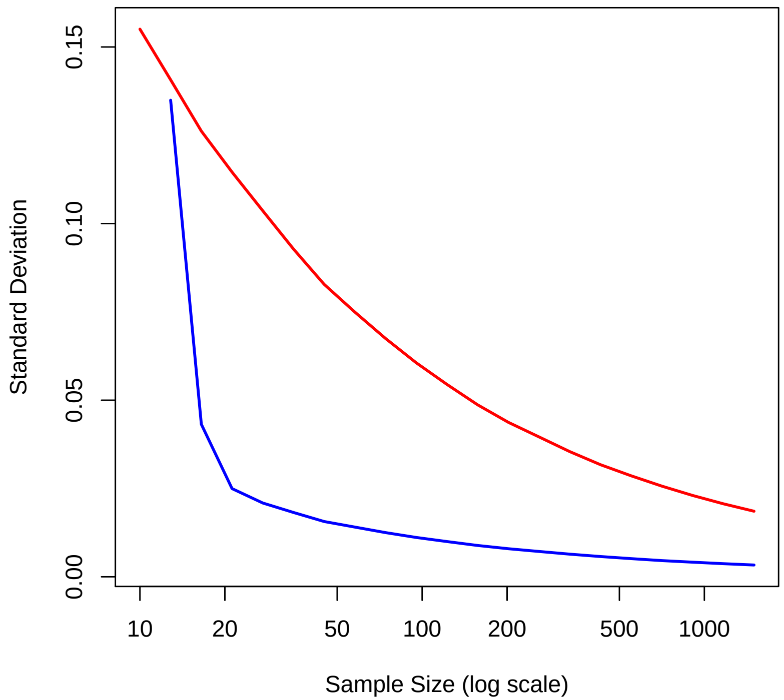
Since $\dfrac{\partial\theta}{\partial\beta_1} = \dfrac{-1}{2\beta_2}$ and $\dfrac{\partial\theta}{\partial\beta_2} = \dfrac{\beta_1}{2\beta_2^2}$, the variance is

$$\frac{1}{4}\begin{bmatrix}\dfrac{-1}{2\beta_2} & \dfrac{\beta_1}{2\beta_2^2}\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \sigma_{12}\\ \sigma_{12} & \sigma_2^2\end{bmatrix}\begin{bmatrix}\dfrac{-1}{2\beta_2} & \dfrac{\beta_1}{2\beta_2^2}\end{bmatrix}^{\mathsf{T}} = \frac{\sigma_1^2\beta_2^2 - 2\beta_1\beta_2\sigma_{12} + \beta_1^2\sigma_2^2}{4\beta_2^2}$$

Estimation of Various Quantifies of Interest

2. Use of Bootstrap

standard deviation of $\widehat{\theta}$,
— delta method vs.
— bootstrap.



Box-Cox Transform

$$y_\lambda = \beta_0 + \beta_1 x + \varepsilon, \text{ with } y_\lambda = \frac{y^{\lambda-1}}{\lambda}$$

with the limiting case $y_0 = \log[y]$.

We assume that for some (unkown) λ_0, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

As in Horowitz (1998) The Bootstrap, use residual bootstrap:

$$y_i^{(b)} = \big(\lambda[\widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{\varepsilon}^{(b)}]\big)^{1/\lambda}$$

Kernel based Regression

Consider some kernel based regression of estimate $m(x)=\mathbb{E}[Y|X=x]$,

$$\widehat{m}_h(x)=\frac{1}{nh\widehat{f}_n(x)}\sum_{i=1}^n y_i k\left(\frac{x-x_i}{h}\right) \text{ where } \widehat{f}_n(x)=\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x-x_i}{h}\right)$$

We have seen that the bias was

$$b_h(x)=\mathbb{E}[\tilde{m}(x)]-m(x)\propto h^2\left(\frac{1}{2}m''(x)+m'(x)\frac{f'(x)}{f(x)}\right)$$

and the variance

$$v_h(x)\propto\frac{\text{Var}[Y|X=x]}{nhf(x)}$$

Further

$$Z_n(x)=\frac{\widehat{m}_{h_n}(x)-m(x)-b_{h_n}(x)}{\sqrt{v_h h_n(x)}}\overset{\mathcal{L}}{\to}\mathcal{N}(0,1)\text{ as }n\to\infty.$$

Kernel based Regression

Idea: convert $Z_n(x)$ into an asymptotically pivotal statistic

Observe that

$$\widehat{m}_h(x) - m(x) \sim \frac{1}{nhf(x)} \sum_{i=1}^{n} [y_i - m(x)] k\left(\frac{x - x_i}{h}\right)$$

so that $v_n(x)$ can be estimated by

$$\widehat{v}_n(x) = \frac{1}{(nh\widehat{f}_n(x))^2} \sum_{i=1}^{n} [y_i - \widehat{m}_h(x)]^2 k\left(\frac{x - x_i}{h}\right)^2$$

then set

$$\widehat{\theta} = \frac{\widehat{m}_h(x) - m(x)}{\sqrt{\widehat{v}_n(x)}}$$

$\widehat{\theta}$ is asymptotically $\mathcal{N}(0,1)$ and it is an asymptotically pivotal statistic

Poisson Regression

Diagnosis period		Reporting-delay interval (quarters):									Total reports to end
Year	Quarter	$0^\dagger$	1	2	3	4	5	6	···	$\geq$14	of 1992
1988	1	31	80	16	9	3	2	8	···	6	174
	2	26	99	27	9	8	11	3	···	3	211
	3	31	95	35	13	18	4	6	···	3	224
	4	36	77	20	26	11	3	8	···	2	205
1989	1	32	92	32	10	12	19	12	···	2	224
	2	15	92	14	27	22	21	12	···	1	219
	3	34	104	29	31	18	8	6	···		253
	4	38	101	34	18	9	15	6	···		233
1990	1	31	124	47	24	11	15	8	···		281
	2	32	132	36	10	9	7	6	···		245
	3	49	107	51	17	15	8	9	···		260
	4	44	153	41	16	11	6	5	···		285
1991	1	41	137	29	33	7	11	6	···		271
	2	56	124	39	14	12	7	10			263
	3	53	175	35	17	13	11				306
	4	63	135	24	23	12					258
1992	1	71	161	48	25						310
	2	95	178	39							318
	3	76	181								273
	4	67									133

Example : see Davison & Hinkley (1997) Bootstrap Methods and Applications, UK AIDS diagnoses, 1988-1992.

Reporting delay can be important

Let j denote year and k denote delay. Assumption

$$N_{j,k} \sim \mathcal{P}(\lambda_{j,k}) \text{ with } \lambda_{j,k} = \exp[\alpha_j + \beta_k]$$

Unreported diagnoses for period j : $\displaystyle\sum_{k \text{ unobserved}} \lambda_{j,k}$

Prediction : $\displaystyle\sum_{k \text{ unobserved}} \widehat{\lambda}_{j,k} = \exp[\widehat{\alpha}_j] \sum_{k \text{ unobserved}} \exp[\widehat{\beta}_k]$

Poisson regression is a GLM : confidence intervals on coefficients are asymptotic.

Let V denote the variance function, then Pearson residuals are

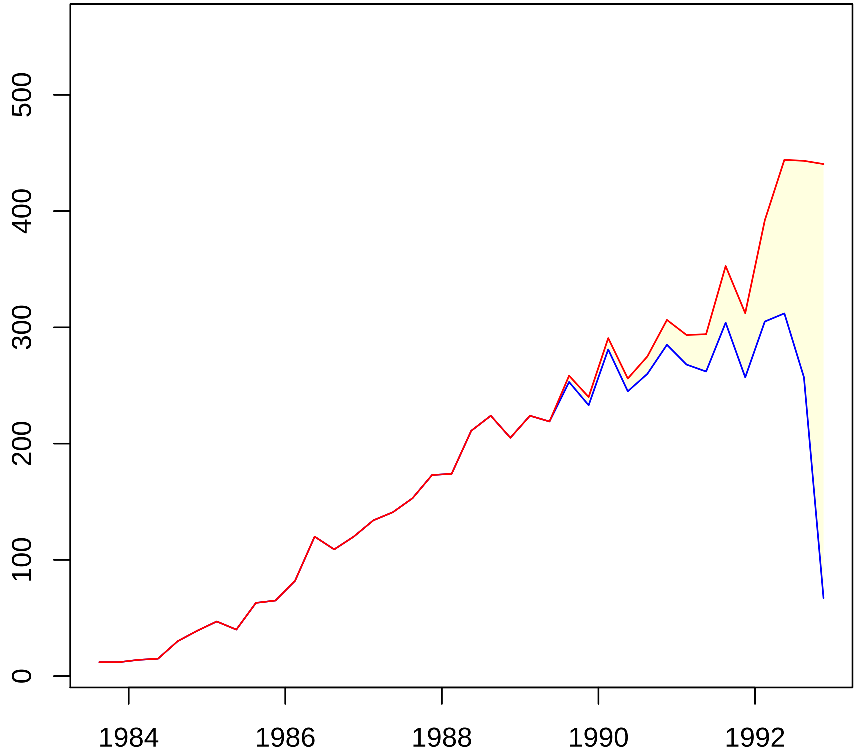
$$\widehat{\epsilon}_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{V[\widehat{\mu}_i]}}$$

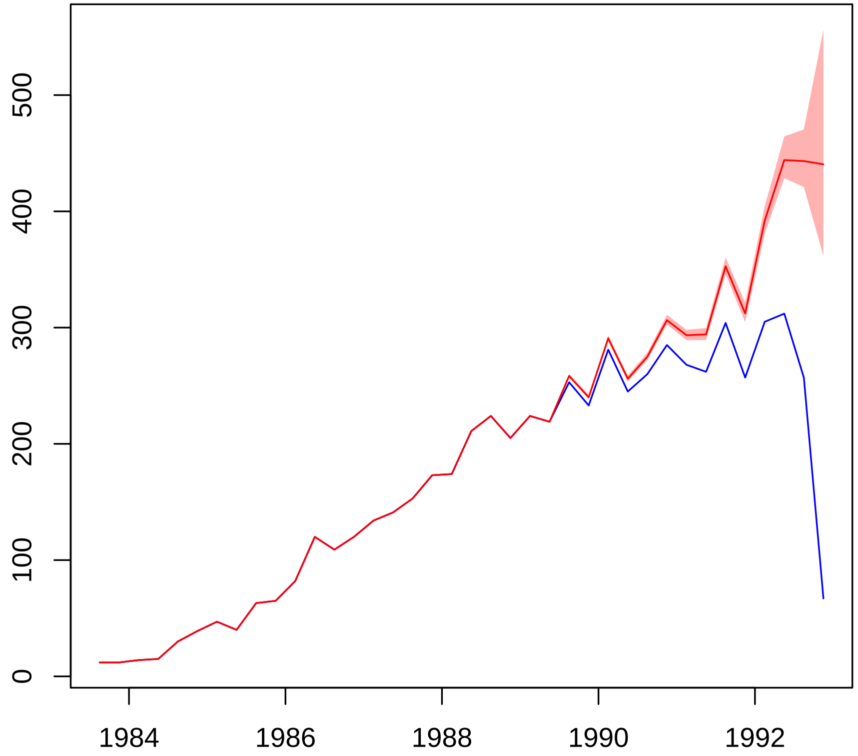
so here

$$\widehat{\epsilon}_{j,k} = \frac{n_{j,k} - \widehat{\lambda}_{j,k}}{\sqrt{\widehat{\lambda}_{j,k}}}$$

Poisson Regression

So bootstrapped responses are $n^\star_{j,k} = \widehat{\lambda}_{j,k} + \sqrt{\widehat{\lambda}_{j,k}} \cdot \widehat{\epsilon}^\star_{j,k}$





Bootstrapping and Logistic Regression

Consider a simple logistic regression,

$$\text{logit}[\pi_i] = \beta_0 + \beta_1 x_i, \text{ i.e. } \pi_i = \frac{\exp[\beta_0 + \beta_1 x_i]}{1 + \exp[\beta_0 + \beta_1 x_i]}.$$

Carpenter & Bithell (2000) suggested a parametric bootstrap

1. Estimate parameters $\widehat{\beta}$

2. Draw boostrap samples $(x_i, y_i^{(b)})$ where $y_i^{(b)} \sim \mathcal{B}(\widehat{\pi}_i)$.

3. Estimate parameters $\widehat{\beta}^{(b)}$

One can also consider a nonparametric bootstrap

Pivotal Case (or not)

In some cases, $G(\cdot, F)$ does not depend on F, $\forall F \in \mathcal{F}$.

Then T_n is said to be pivotal, relative to $\mathcal{F}$.

Example : consider the case of Gaussian residuals, $\mathcal{F} = \mathcal{F}_{\text{gaussian}}$. Then

$$T = \frac{\overline{y} - \mathbb{E}[Y]}{\widehat{\sigma}} \sim \mathcal{S}td(n-1)$$

which does not depend on F (but it does depend on $\mathcal{F}$)

If T_n is not pivotal, it is still possible to look for bounds on $G_n(t, F)$,

$$B_n(t) = \left[\inf_{F \in \mathcal{F}_\star} \{G_n(t, F)\}; \sup_{F \in \mathcal{F}_\star} \{G_n(t, F)\}\right]$$

for instance, when a set of *reasonable values* for $\mathcal{F}_\star$ is provided, by an expert.

Pivotal Case (or not)

$$B_n(t) = \left[\inf_{F\in\mathcal{F}_\star}\{G_n(t,F)\};\ \sup_{F\in\mathcal{F}_\star}\{G_n(t,F)\}\right]$$

In the parametric case, set

$$\mathcal{F}_\star = \{F_\theta, \theta \in IC\}$$

where IC is some confidence interval.

In the nonparametric case, use Kolomogorov-Smirnov statistics to get bounds, using quantiles of

$$\sqrt{n}\sup\{|\widehat{F}_n(t) - F_0(t)|\}$$

Pivotal Function and Studentized Statistics

It is interesting to studentize any statistics.

Let v denote the variance of $\widehat{\theta}$ (computed using $\{y_1,\cdots,y_n\}$. Then set

$$Z=\frac{\widehat{\theta}-\theta}{\sqrt{v}}$$

If quantiles of Z are known (and denoted z_α), then

$$\mathbb{P}\Big(\widehat{\theta}+\sqrt{v}z_{\alpha/2}\leq\theta\leq\widehat{\theta}+\sqrt{v}z_{1-\alpha/2}\Big)=1-\alpha$$

Idea : use a (double) boostrap procedure

Pivotal Function and Double Bootstrap Procedure

1. Generate a bootstrap sample $\boldsymbol{y}^{(b)} = \{y_1^{(b)}, \cdots, y_n^{(b)}\}$
2. Compute $\widehat{\theta}^{(b)}$
3. From $\boldsymbol{y}^{(b)}$ generate β bootstrap sample, and compute $\{\widehat{\theta}_1^{(b)}, \cdots, \widehat{\theta}_\beta^{(b)}\}$
4. Compute $\widehat{v}^{(b)} = \dfrac{1}{\beta}\displaystyle\sum_{j=1}^{\beta}\big(\widehat{\theta}_j^{(b)} - \overline{\theta}^{(b)}\big)^2$
5. Set $z^{(b)} = \dfrac{\widehat{\theta}^{(b)} - \widehat{\theta}}{\sqrt{\widehat{v}^{(b)}}}$

Then use $\{z^{(1)}, \cdots, z^{(B)}\}$ to estimate the distribution of z's (and some quantiles).

$$\mathbb{P}\left(\widehat{\theta} + \sqrt{v}z_{\alpha/2}^{(B)} \leq \theta \leq \widehat{\theta} + \sqrt{v}z_{1-\alpha/2}^{(B)}\right) = 1 - \alpha$$

Why should we studentize ?

Here $Z \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ as $n \to \infty$ (CLT). Using Edgeworth series,

$$\mathbb{P}[Z \leq z|F] = \Phi(z) + n^{-1/2}p(z)\varphi(z) + O(n^{-1})$$

for some quadratic polynomial $p(\cdot)$. For $Z^{(b)}$

$$\mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = \Phi(z) + n^{-1/2}\widehat{p}(z)\varphi(z) + O(n^{-1})$$

where $\widehat{p}(z) = p(z) + O(n^{-1/2})$, so

$$\mathbb{P}[Z \leq z|F] - \mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = O(n^{-1})$$

But if we do not studentize, $Z = (\widehat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,\nu)$ as $n \to \infty$ (CLT).

Why should we studentize ?

Using Edgeworth series,

$$\mathbb{P}[Z \leq z|F] = \Phi\left(\frac{z}{\sqrt{\nu}}\right) + n^{-1/2}p'\left(\frac{z}{\sqrt{\nu}}\right)\varphi\left(\frac{z}{\sqrt{\nu}}\right) + O(n^{-1})$$

for some quadratic polynomial $p(\cdot)$. For $Z^{(b)}$

$$\mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = \Phi\left(\frac{z}{\sqrt{\widehat{\nu}}}\right) + n^{-1/2}\widehat{p}'\left(\frac{z}{\sqrt{\widehat{\nu}}}\right)\varphi\left(\frac{z}{\sqrt{\widehat{\nu}}}\right) + O(n^{-1})$$

recall that $\widehat{\nu} = \nu + 0(n^{-1/2})$, and thus

$$\mathbb{P}[Z \leq z|F] - \mathbb{P}[Z^{(b)} \leq z|\widehat{F}] = O(n^{-1/2})$$

Hence, studentization reduces error, from $O(n^{-1/2})$ to $O(n^{-1})$

Variance estimation

The estimation of $\text{Var}[\widehat{\theta}]$ is necessary for studentized boostrap.

- double bootstrap (used here)
- delta method
- jackknife (leave-one-out)

Double Bootstrap

Requieres $B \times \beta$ resamples, e.g. $B \sim 1,000$ while $\beta \sim 100$

Delta Method

Let $\widehat{\tau} = g(\widehat{\theta})$, with $g'(\theta) \neq 0$.

$$\mathbb{E}[\widehat{\tau}] = g(\theta) + O(n^{-1})$$

$$\text{Var}[\widehat{\tau}] = \text{Var}[\widehat{\theta}]g'(\theta)^2 + O(n^{-3/2})$$

Variance estimation

Idea: find a transformation such that $\text{Var}[\widehat{\tau}]$ is constant. Then

$$\text{Var}[\widehat{\theta}] \sim \frac{\text{Var}[\widehat{\tau}]}{g'(\widehat{\theta})^2}$$

There is also a nonparametric delta method, based on the influence function.

Influence Function and Taylor Expansion

Taylor expansion

$$t(y) = t(x) + \int_x^y f'(z)cdz\ t(x) + (y-x)f'(x)$$

$$t(G) = t(F) + \int_{\mathbb{R}} L_t(z,F)dG(z)$$

where L_t is the Fréchet derivative,

$$L_t(z,F) = \left.\frac{\partial[(1-\epsilon)F + \epsilon\Delta_z]}{\partial\epsilon}\right|_{\epsilon=0}$$

where $\Delta_z(t) = \mathbf{1}(t > z)$ denote the cdf of the Dirac measure in z.

For instance, observe that

$$t(\widehat{F}_n) = t(F) + \frac{1}{n}\sum_{i=1}^n L_t(y_i,F)$$

Influence Function and Taylor Expansion

This can be used to estimate the variance. Set

$$V_L = \frac{1}{n^2}\sum_{i=1}^n L(y_i,F)^2$$

where $L(y,F)$ is the influence function for $\theta=t(F)$ for observation at y when distribution is F.

The empirical version is $\ell_i = L(y_i,\widehat{F})$ and set

$$\widehat{V}_L = \frac{1}{n^2}\sum_{i=1}^n \ell_i^2$$

Example : let $\theta=\mathbb{E}[X]$ with $X\sim F$, then

$$\widehat{\theta}=\overline{y}_n=\sum_{i=1}^n \frac{1}{n}y_i=\sum_{i=1}^n \omega_i y_i \text{ where } \omega_i=\frac{1}{n}$$

Influence Function and Taylor Expansion

Change ω's in direction j :

$$\omega_j = \epsilon + \frac{1-\epsilon}{n}, \text{ while } \forall i \neq j, \omega_i = \frac{1-\epsilon}{n},$$

then $\widehat{\theta}$ changes in

$$\underbrace{[y_h - \widehat{\theta}]}_{\ell_j}\epsilon + \widehat{\theta}$$

Hence, ℓ_j is the standardized chance in $\widehat{\theta}$ with an increase in direction j, and

$$\widehat{V}_L = \frac{n-1}{n}\frac{\text{Var}[X]}{n}.$$

Example : consider a ratio, $\theta = \dfrac{\mathbb{E}[X]}{\mathbb{E}[Y]}$, then

$$\widehat{\theta} = \frac{\overline{x}_n}{\overline{y}_n} \text{ and } \ell_j = \frac{x_j - \widehat{\theta} y_j}{\overline{y}_n}$$

Influence Function and Taylor Expansion

so that

$$\widehat{V}_L = \frac{1}{n^2}\sum_{i=1}^{n}\left(\frac{x_j - \widehat{\theta} y_j}{\overline{y}_n}\right)^2$$

Example : consider a correlation coefficient,

$$\theta = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]}{\sqrt{\left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right)\cdot\left(\mathbb{E}[Y^2] - \mathbb{E}[Y]^2\right)}}$$

Let $\overline{xy} = n^{-1}\sum x_i y_i$, so that

$$\widehat{\theta} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\sqrt{\left(\overline{x^2} - \overline{x}^2\right)\cdot\left(\overline{y^2} - \overline{y}^2\right)}}$$

Jackknife

An approximation of ℓ_i is $\ell_i^\star = (n-1)\big(\widehat{\theta} - \widehat{\theta}_{(-j)}\big)$ where $\widehat{\theta}_{(-j)}$ is the statistics computed from sample $\{y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n\}$

One can define Jackknife bias and Jackknife variance

$$b^\star = \frac{-1}{n}\sum_{i=1}^n \ell_i^\star \text{ and } v^\star = \frac{1}{n(n-1)}\left(\sum_{i=1}^n \ell_i^{\star 2} - nb^{\star 2}\right)$$

cf numerical differentiation when $\epsilon = -\dfrac{1}{(n-1)}$.

Convergence

Given a sample $\{y_1, \cdots, y_n\}$, i.id. with distribution F, set

$$\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(y_i \le t)$$

Then

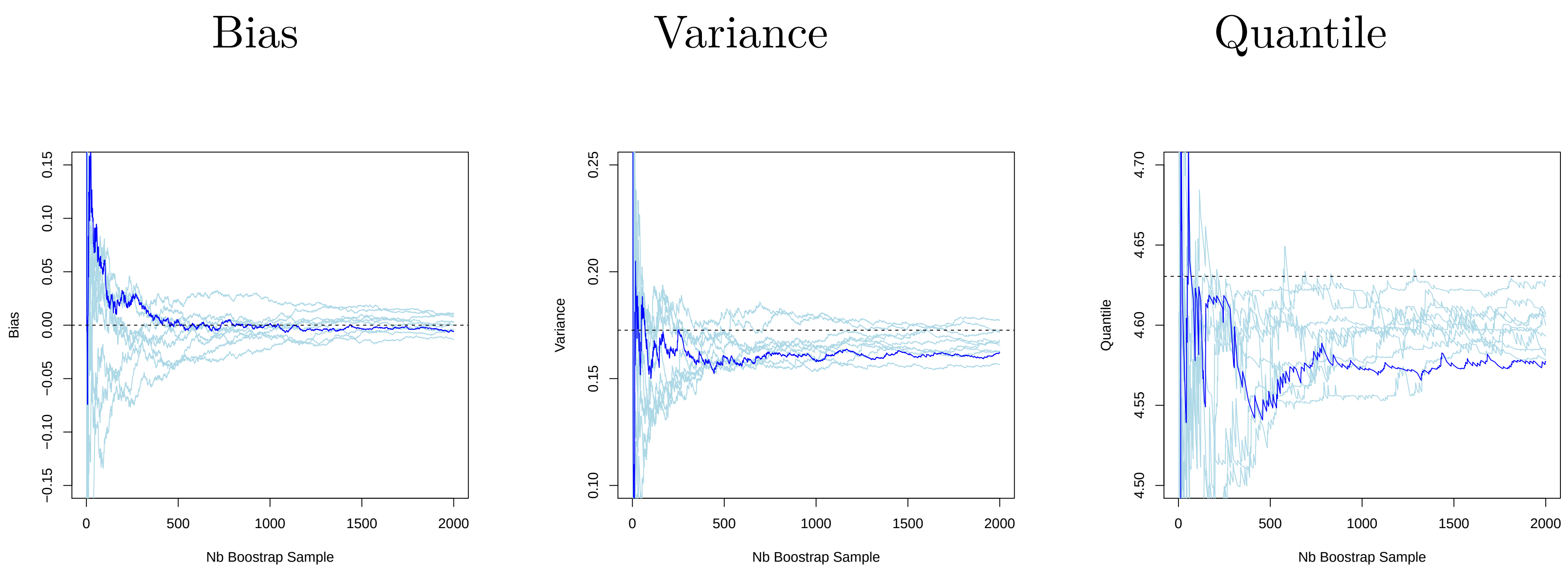
$$\sup\left\{|\widehat{F}_n(t) - F_0(t)|\right\} \xrightarrow{\mathbb{P}} 0, \text{ as } n \to \infty.$$

How many Boostrap Samples?

Easy to take $B \geq 5000$

$R > 100$ to estimate bias or variance

$R > 1000$ to estimate quantiles



How many Boostrap Samples?

```
> reg=lm(dist~speed,data=cars)
> E=residuals(reg)
> Y=predict(reg)
> beta1=rep(NA,2000)
> MB=MV=MQ=matrix(NA,10,2000)
> for(i in 1:10){
+ BIAS=VAR=QUANT=beta1
+ for(b in 1:2000){
+ carsS=data.frame(speed=cars$
    speed,
+ dist=Y+sample(E,size=50,replace
    =TRUE))
```

```
+ beta1[b]=lm(dist~speed,data=
    carsS)$coefficients[2]
+ BIAS[b]=mean(beta1[1:b])-reg$
    coefficients[2]
+ VAR[b]=var(beta1[1:b])
+ QUANT[b]=quantile(beta1[1:b
    ],.95)
+ }
+ MB[i,]=BIAS
+ MV[i,]=VAR
+ MQ[i,]=QUANT
```

Consistency

We expect something like

$$G_n(t,\widehat{F}_n) \sim G_\infty(t,\widehat{F}_n) \sim G_\infty(t,F_0) \sim G_n(t,F_0)$$

$G_n(t,\widehat{F}_n)$ is said to be consistent if under each $F_0 \in \mathcal{F}$,

$$\sup_t \in \mathbb{R}\big\{|G_n(t,\widehat{F}_n) - G_\infty(t,F_0)|\big\} \overset{\mathbb{P}}{\to} 0$$

Example: let $\theta = \mathbb{E}_{F_0}(X)$ and consider $T_n = \sqrt{n}(\overline{X} - \theta)$. Here

$$G_n(t,F_0) = \mathbb{P}_{F_0}(T_n \leq t)$$

Based on boostrap samples, a bootstrap version of T_n is

$$T_n^{(b)} = \sqrt{n}\big(\overline{X}^{(b)} - \overline{X}\big) \text{ since } \overline{X} = \mathbb{E}_{\widehat{F}_n}(X)$$

and $G_n(t,\widehat{F}_n) = \mathbb{P}_{\widehat{F}_n}(T_n \leq t)$

Consistency

Consider a regression model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$

The natural assumption is $\mathbb{E}[\varepsilon_i | \boldsymbol{X}] = 0$ with ε_i's i.id.$\sim F$.

The parameter of interest is $\theta = \beta_j$, and let $\widehat{\beta}_j = \theta(\widehat{F}_n)$.

1. The statistics of interest is $T_n = \sqrt{n}\big[\widehat{\beta}_j - \beta_j\big]$.

We want to know $G_n(t, F_0) = \mathbb{P}_{F_0}(T_n \leq t)$.

Let $\boldsymbol{x}^{(b)}$ denote a bootsrap sample.

Compute $T_n^{(b)} = \sqrt{n}\big(\widehat{\beta}_j^{(b)} - \widehat{\beta}_j\big)$, and then

$$G_n(t, F_n) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(T_n^{(b)} \leq t)$$

Consistency

2. The statistics of interest is $T_n = \sqrt{n}\dfrac{[\widehat{\beta}_j - \beta_j]}{\sqrt{\text{Var}[\widehat{\beta}_j]}}$.

We want to know $G_n(t, F_0) = \mathbb{P}_{F_0}(T_n \leq t)$.

Let $\boldsymbol{x}^{(b)}$ denote a bootsrap sample.

Compute $T_n^{(b)} = \sqrt{n}\dfrac{[\widehat{\beta}_j^{(b)} - \widehat{\beta}_j]}{\sqrt{\text{Var}^{(b)}[\widehat{\beta}_j]}}$, and then

$$G_n(t, F_n) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(T_n^{(b)} \leq t)$$

This second option is more accurate than the first one :

Consistency

The approximation error of bootstrap applied to asymptotically pivotal statistic is smaller than the approximation error of bootstrap applied on asymptotically non-pivotal statistic, see Horowitz (1998) The Bootstrap.

Here, asymptotically pivotal means that

$$G_\infty(t, F) = G_\infty(t),\ \forall F \in \mathcal{F}.$$

Assume now that the quantity of interest is $\theta = \text{Var}[\widehat{\boldsymbol{\beta}}]$.

Consider a bootstrap procedure, then one can prove that

$$\underset{B,n\rightarrow\infty}{\text{plim}} \left\{ \frac{1}{B} \sum_{b=1}^{B} \sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\boldsymbol{\beta}}\big)\sqrt{n}\big(\widehat{\boldsymbol{\beta}}^{(b)} - \widehat{\boldsymbol{\beta}}\big)^{\mathsf{T}} \right\}$$

$$= \underset{n\rightarrow\infty}{\text{plim}} \left\{ n\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big)\big(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\big)^{\mathsf{T}} \right\}$$

More on Testing Procedures

Consider a sample $\{y_1, \cdots, y_n\}$. We want to test some hypothesis H_0. Consider some test statistic $t(\boldsymbol{y})$

Idea: t takes large values when H_0 is not satisfied.

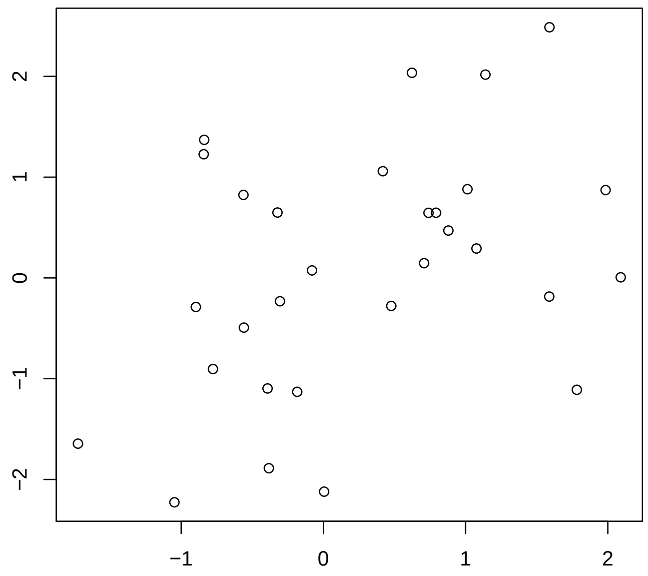
The p-value is $p = \mathbb{P}[T > t_{\text{obs}}|H_0]$.

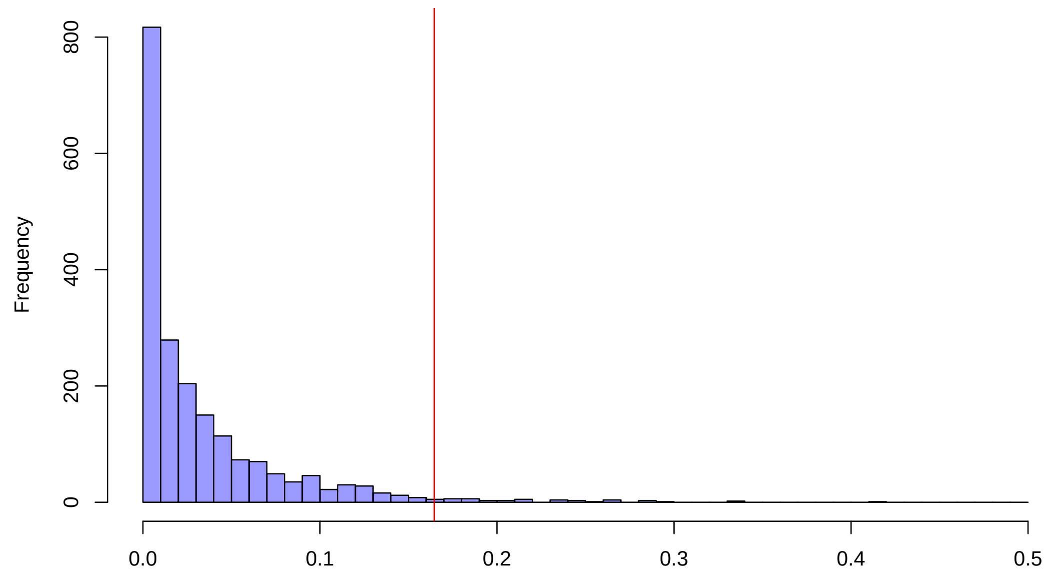
Boostrap/simulations can be used to estimate p, by simulation from H_0.

1. Generate $\boldsymbol{y}^{(s)} = \{y_1^{(s)}, \cdots, y_n^{(s)}\}$ generated from H_0.
2. Compute $t^{(s)} = t(\boldsymbol{y}^{(s)})$
3. Set $\displaystyle\widehat{p} = \frac{1}{1+S}\left(1 + \sum_{s=1}^{S} \mathbf{1}(t^{(s)} \geq t_{\text{obs}})\right)$

Example : testing independence, let t denote the square of the correlation coefficient.

Under H_0 variables are independent, so we can bootstrap independently x's and y's.





With this bootstap procedure, we estimate

$$\widehat{p} = \mathbb{P}\big(T \geq t_{\text{obs}} | \widehat{H}_0\big)$$

which is not the same as

$$p = \mathbb{P}\big(T \geq t_{\text{obs}} | H_0\big)$$

More on Testing Procedures

In a parametric model, it can be interesting to use a sufficient statistic W. One can prove that

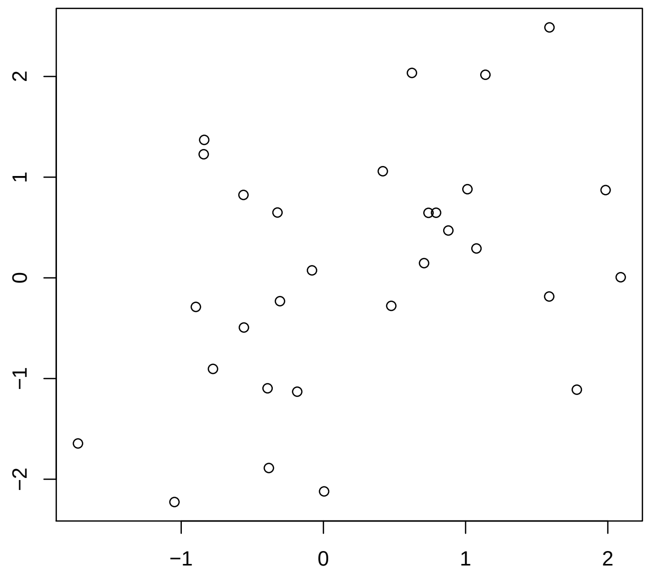
$$p = \mathbb{P}\big(T \geq t_{\text{obs}}|\widehat{H}_0, W\big)$$

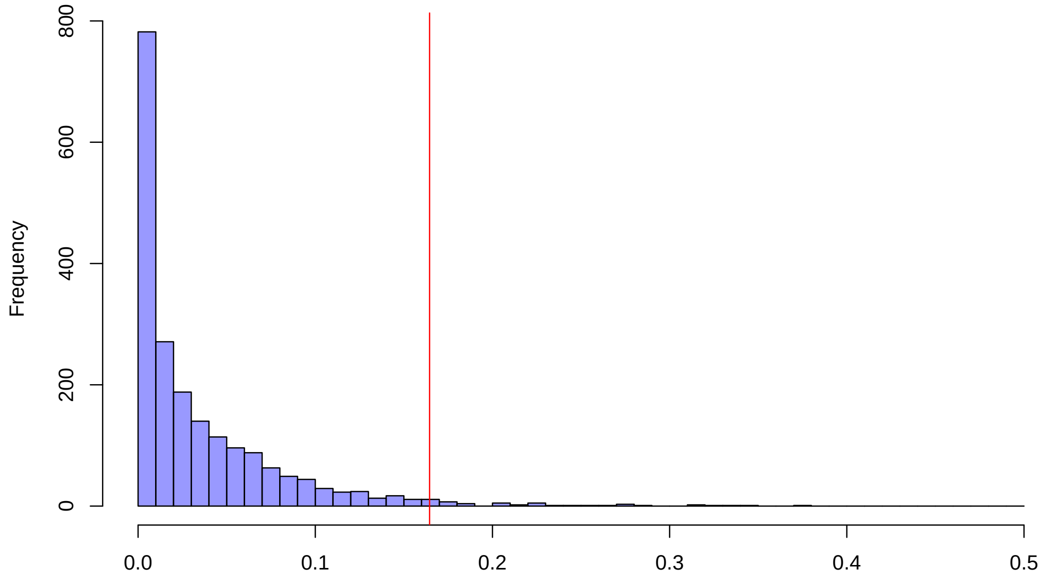
The problem is to generate from this conditional distribution...

Example : for the independence test, we should sample from $\widehat{F}_x$ and $\widehat{F}_y$ with fixed margins.

Bootstrap should be here without replacement.

More on Testing Procedures





More on Testing Procedures

But this nonparametric bootstrap fails when the Gaussian Central Limit Theorem does not apply (Mammen's theorem)

Example $X \sim$ Cauchy : limit distribution $G_\infty(t, F)$ is not continuous, in F

Example : distribution of the maximum of the support (see Bickel & Freedman (1981)): $X \sim \mathcal{U}([0, \theta_0])$

$T_n = n(\theta_n - \theta_0)$ with $\theta_n = \max\{X_1, \cdots, X_n\}$

Set $T_n^{(b)} = n(\theta_n^{(b)} - \theta_n)$, and $\theta_n^{(b)} = \max\{X_1^{(b)}, \cdots, X_n^{(b)}\}$

Here $T_n \xrightarrow{\mathcal{L}} \mathcal{E}(1)$, exponential distribution, but not $T_n^{(b)}$, since $T_n^{(b)} \geq 0$ (we juste resample), and

$$\mathbb{P}[T_n^{(b)} = 0] = 1 - \mathbb{P}[T_n^{(b)} > 0] = 1 - \left(1 - \frac{1}{n}\right)^n \sim 1 - e^{-1}.$$

Resampling or Subsampling ?

Why not draw subsamples of size $m < n$?

- with replacement, see m out of n boostrap
- without replacement, see subsampling boostrap

Less accurate than bootstrap when bootstrap works... but might work when bootstrap does not work

Exemple : maximum of the support, $Y_i \sim \mathcal{U}([0,\theta])$,

$$\mathbb{P}_{\widehat{F}_n}[T_m^{(b)} = 0] = 1 - \left(1 - \frac{1}{n}\right)^m \sim 1 - e^{-m/n} \sim 0$$

if $m = o(n)$.

From Bootstrap to Bagging

Bagging was introduced in Breiman (1996) Bagging predictors

1. sample a boostrap sample $(y_i^{(b)}, \boldsymbol{x}_i^{(b)})$ by resampling pairs
2. estimate a model $\widehat{m}^{(b)}(\cdot)$

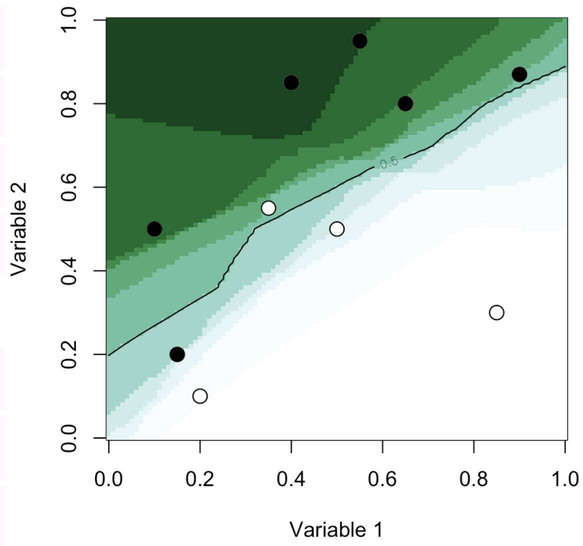
The bagged estimate for m is then $m_{\text{bag}}(\boldsymbol{x}) = \dfrac{1}{B}\sum_{b=1}^{B}\widehat{m}^{(b)}(\boldsymbol{x})$

In the context of a logistic regression, draw pairs $(y_i, \boldsymbol{x}_i)$ randomly, with probabilities $1/n$,

From Bootstrap to Bagging

```
L_logit = list()
n = nrow(df)
for(s in 1:1000){
  df_s = df[sample(1:n,size=n,replace=TRUE),]
  L_logit[[s]] = glm(y~., df_s, family=binomial
    )}
```

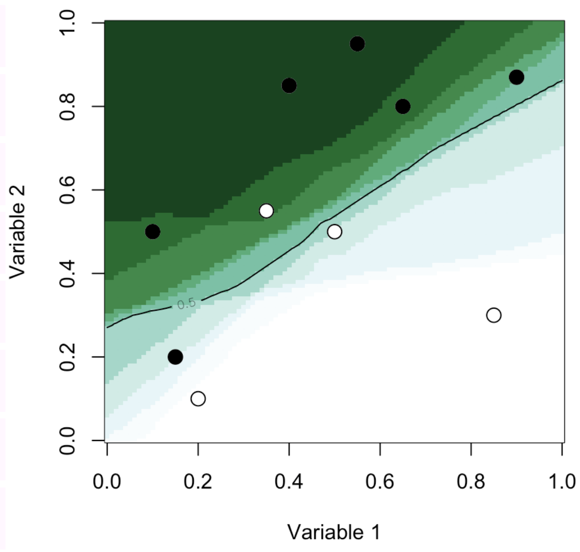
```
  p = function(x){
  nd = data.frame(x1=x[1], x2=x[2])
  unlist(lapply(1:1000,function(z)
  predict(L_logit[[z]],newdata=nd,type="
    response")))}
```



An alternative is to generate a fake sample. Keep all $\boldsymbol{x}_i$'s and for each of them, draw $Y_{i,b} \sim \mathcal{B}(\widehat{m}_S(\boldsymbol{x}_i))$ where $\widehat{m}(\boldsymbol{x}) = \mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}].m(\boldsymbol{x}) = P[Y = 1|\boldsymbol{X} = \boldsymbol{x}]$.

From Bootstrap to Bagging

```
L_logit = list()
n = nrow(df)
reg = glm(y~x1+x2, df, family=binomial)
for(s in 1:100){
  df_s = df
  df_s$y = factor(rbinom(n,size=1,prob=predict(reg,type="response")),labels=0:1)
  L_logit[[s]] = glm(y~., df_s, family=binomial)
}
```



From Bootstrap to Bagging

Random forests are obtained using

```r
L_tree = list()
for(s in 1:1000){
  idx = sample(1:n, size=n, replace=TRUE)
  L_tree[[s]] = rpart(as.factor(PRONO)~.,
    myocarde[idx,])
}
```

Actually, random forest are slightly different : at each node, variables are randomly selected.