El método *bootstrap* para correlaciones y regresión lineal

El t-test para correlaciones que vimos antes requiere que ambas series sigan una distribución normal, por más que se suele usar sin verificar .

El método bootstrap se puede usar para hacer tests de significancia en esos casos, y también para hacer estimaciones alternativas de parámetros (correlaciones, coeficientes de regresión, etc).

El bootstrap es un método de re-

Bootstrap: re-muestreo de los datos originales **con reemplazo** (o sea que el mismo dato puede aparecer más de una vez en la serie remuestreada).

– Dada una serie de longitud n, se toma un número arbitrario ("grande") de muestras de longitud n cuyos elementos son
tomados al azar.

– La estadística de las nuevas muestras provee más información sobre la población.

(El jackknife va suprimiendo un elemento por vez, obteniendo así n muestras
de longitud n-1.)

Sean: t=temp. mínima mensual en julio
pp=precip. ac. mensual en noviembre

r=corr(pp,t)

r = 0.199

Haciendo un test de bondad de ajuste, p. ej., el chi-cuadrado (chi2gof) para t y pp,

obtenemos que la hipótesis de gausianidad no se rechaza al 5% para t, pero sí para pp.

De todos modos, calculamos el estadístico t para la hipótesis H0: r=0

$$T = |(r)| \frac{\sqrt{(n-2)}}{\sqrt{(1-r^2)}}$$

tcalc = r * ((length(t)-2)/(1-r^2))^0.5

tcalc = 1.475

tcrit = tinv(0.975,length(t)-2) % 0.05 y test de 2 extremos

tcrit = 2.006

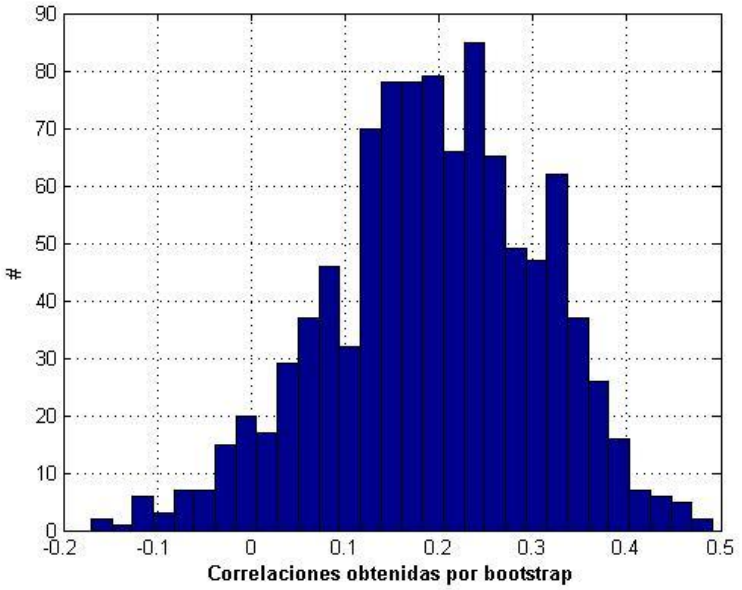
Una alternativa para calcular la significancia es el bootstrap.

```
rcorr_1000=bootstrp(1000,'corr',t,pp);
% remuestrea los datos originales 1000 veces y calcula la correlación para cada una d

hist(rcorr_1000,30)

mean(rcorr_1000)
ans = 0.194

std(rcorr_1000)
ans = 0.113
```



El valor medio es muy similar al de la muestra, pero la dispersión es relativamente grande.

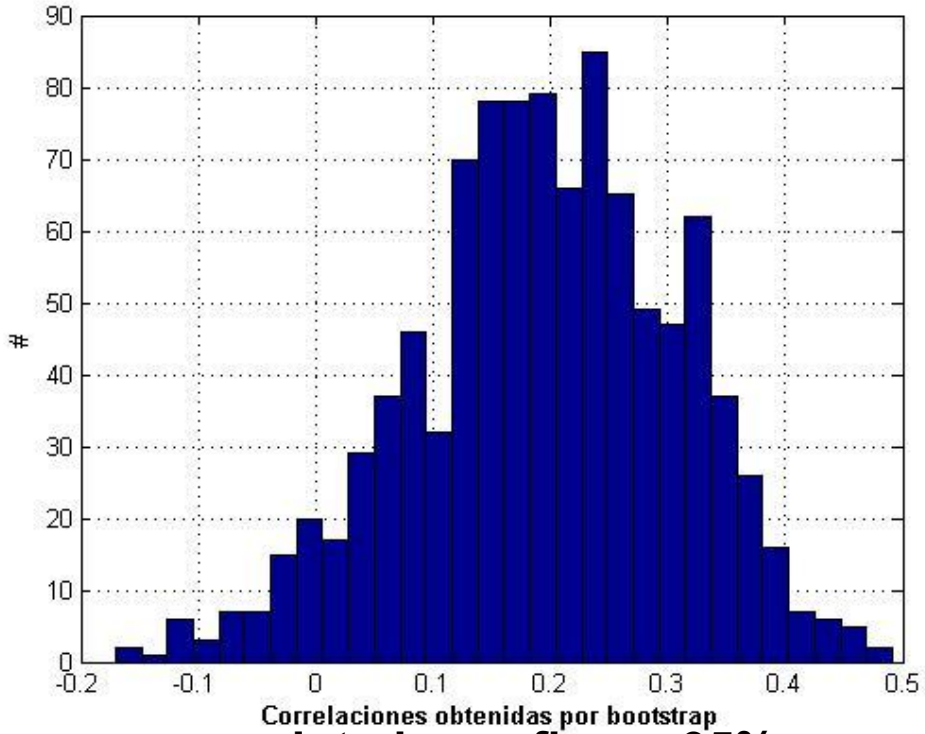
```
corr_fu= @(x,y) corr(x,y);

bootci(2000,{corr_fu,t,p})

ans =

-0.0250
 0.4200
```

El intervalo de 95% de confianza contiene al 0, por lo que no rechazamos la hipótesis nula r=0



Int. de confianza 95%

El bootstrap tambien permite detectar outliers.

scatter(t,pp,’r’,’filled’)

La corr de Pearson es sensible a los
outliers.
Agreguemos un valor atípico:

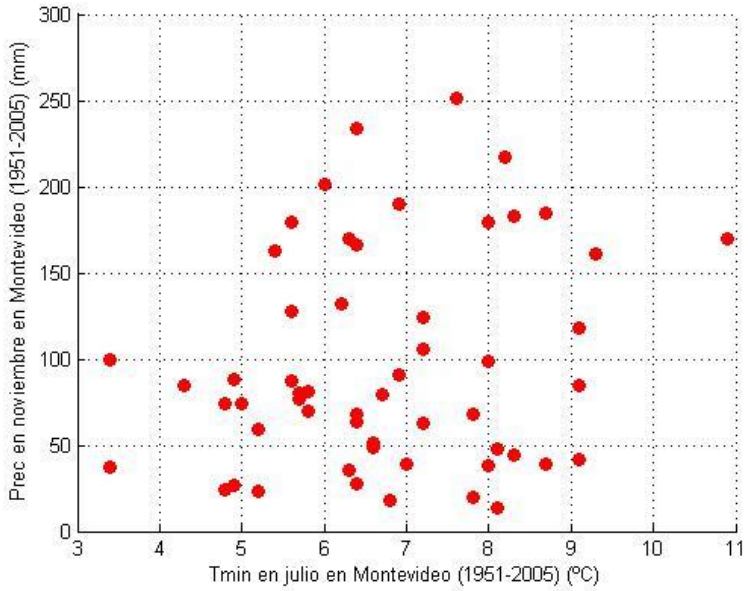
t(56)=50;
pp(56)=1000;

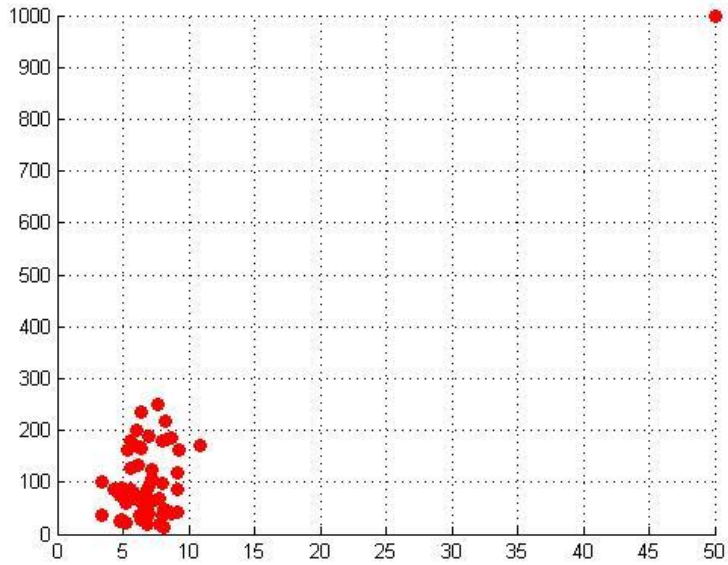
scatter(t,pp,'r','filled'), grid

Ahora:
corr(pp,t)

ans = 0.883

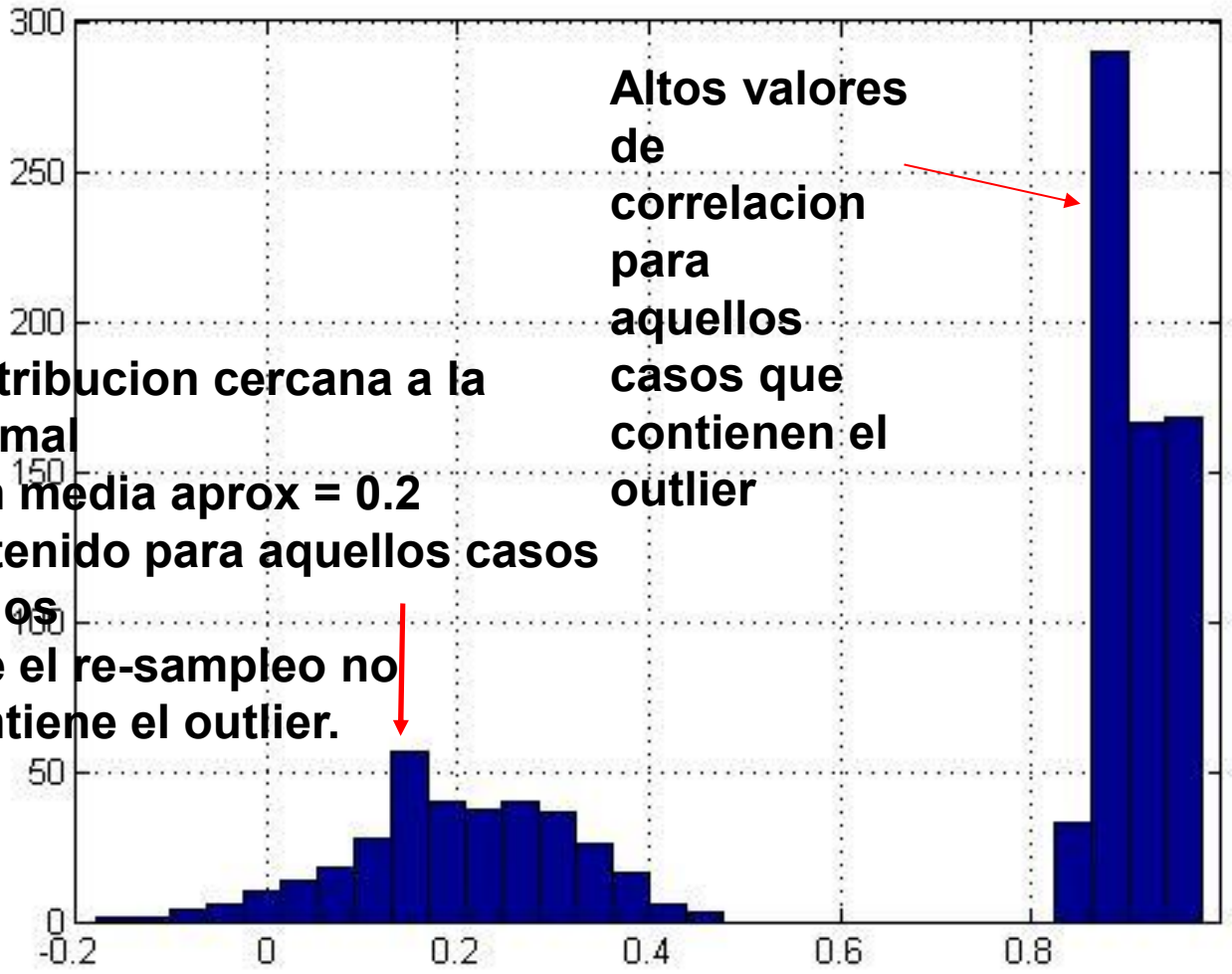
(el gran aumento es debido al outlier)





Volvemos a hacer el bootstrap:

rcorr_1000=bootstrp(1000,'corr',t,pp);



Bootstrap y regresion lineal

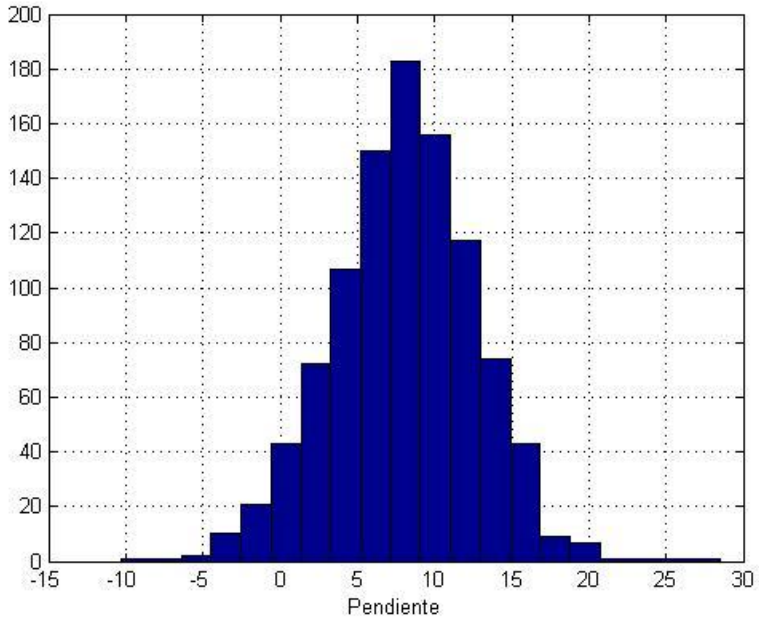
Es posible usar la técnica de bootstrap para estimar los coeficientes de regresión y un intervalo de confianza.
Seguimos con el mismo ejemplo. Sabemos que el ajuste lineal no va a ser bueno, debido a la baja correlación.

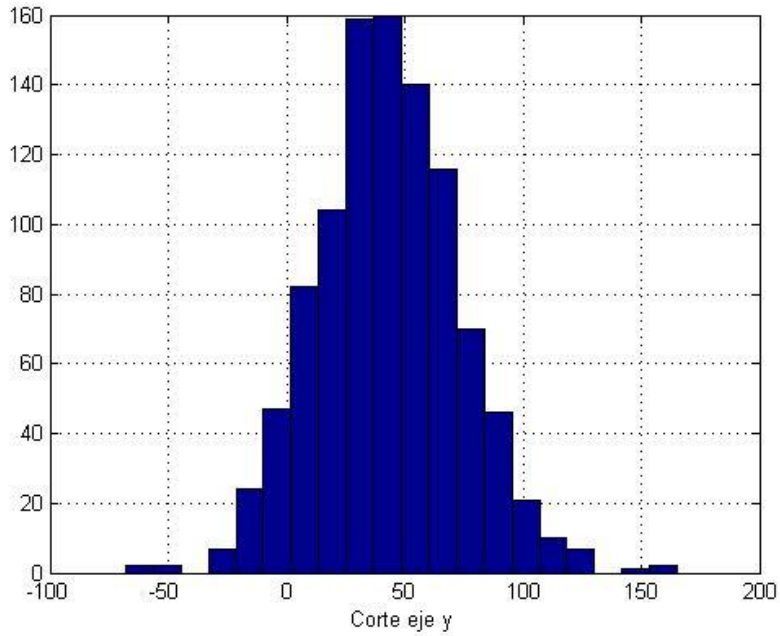
```
coef=polyfit(t,pp,1)

coef =

 8.0223   42.8574

coef_boostrp=bootstrp(1000,'polyfit',t,pp,1); % obtenemos 1000 pares
de coeficientes (pendiente y término independiente)
```





hist(coef_boostrp(:,1),20)
hist(coef_boostrp(:,2),20)

mean(coef_boostrp(:,1))
ans = 8.0419
std(coef_boostrp(:,1))
ans =4.6494

mean(coef_boostrp(:,2))
ans = 42.8546
std(coef_boostrp(:,2))
ans = 30.0928

Los intervalos de confianza de 95%:

regre=@(t,pp) polyfit(t,pp,1);
bootci(2000,{regre,t,pp})

ans =

-2.2197	**-13.4153**
16.7516	**111.4381**

El 0 pertenece al IC de la pendiente