Bootstrapping a Ratio Statistic

using the boot package in R

Solon Karapanagiotis

6 August 2016

The `boot` package needs to be installed and loaded.

```
library(boot)
```

We consider the population of the most populated cities in US in 1920 and 1930. The dataset contains n=49 data pairs, each corresponding to a US city, the pair being the populations of the city (in 1000's), which we denote by u and x.

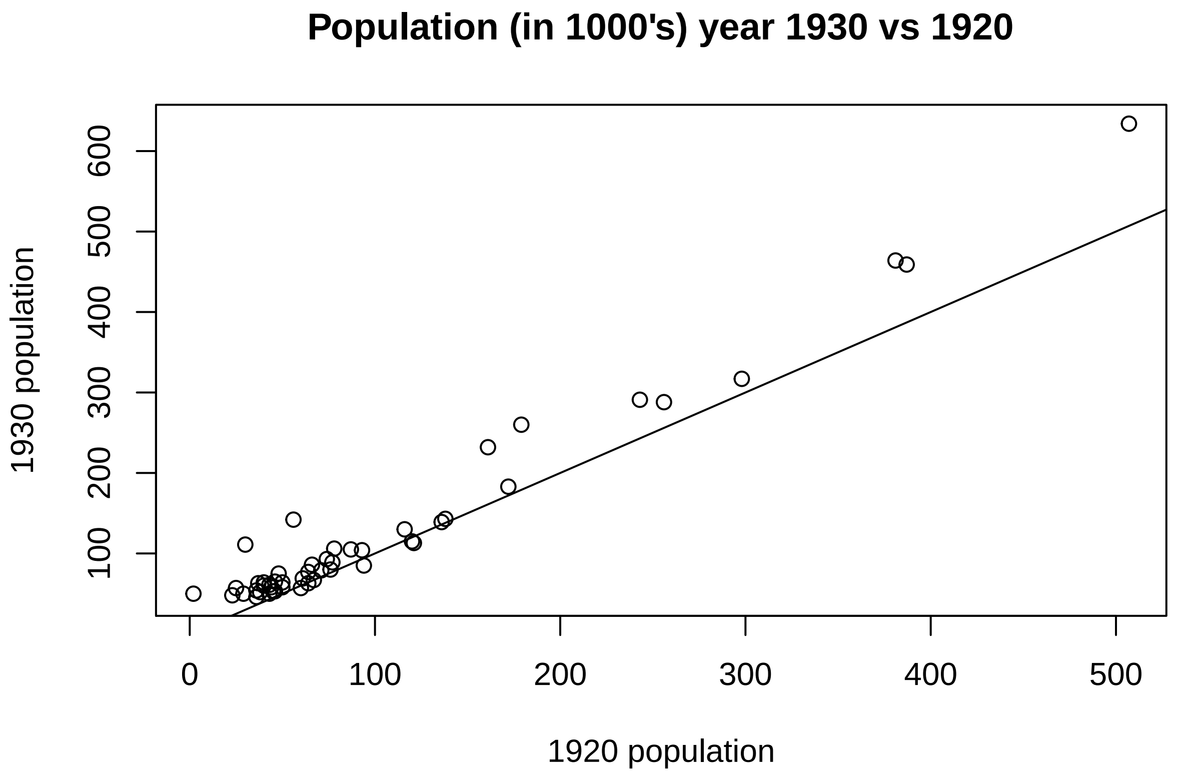
```
str(bigcity)
```

```
## 'data.frame':    49 obs. of  2 variables:
##  $ u: num  138 93 61 179 48 37 29 23 30 2 ...
##  $ x: num  143 104 69 260 75 63 50 48 111 50 ...
```

- `u` is the 1920 population.
- `x` is the 1930 population.

A possible research question is if there is a difference in the population from 1920 to 1930. The scatter plot suggests this is true since the majority of the points lie away from the 45 degree line.

```
plot(bigcity$u, bigcity$x, xlab="1920 population", ylab="1930 population", main="Population (in 1000's)
abline(0,1)
```



In order to answer the research question we use the ratio of expectations

$$\theta = \frac{E(X)}{E(U)}$$

This ratio is the parameter of interest. And the hypotheses are

$$H_0 : \theta = 1,$$

$$H_a : \theta \neq 1$$

H_0 corresponds to no difference in the population between 1920 and 1930. A natural (but biased) estimator of θ is

$$\widehat{\theta} = \frac{\bar{x}}{\bar{u}}$$

with $\bar{x} = \sum\limits_{b=1}^{49} \frac{x_b}{n}$ and $\bar{u} = \sum\limits_{b=1}^{49} \frac{u_b}{n}$.

The observed value is $\widehat{\theta} = 1.239$, indicating higher population the year 1930, on average.

```r
(theta_hat <- mean(bigcity$x)/mean(bigcity$u))
```

```
## [1] 1.239019
```

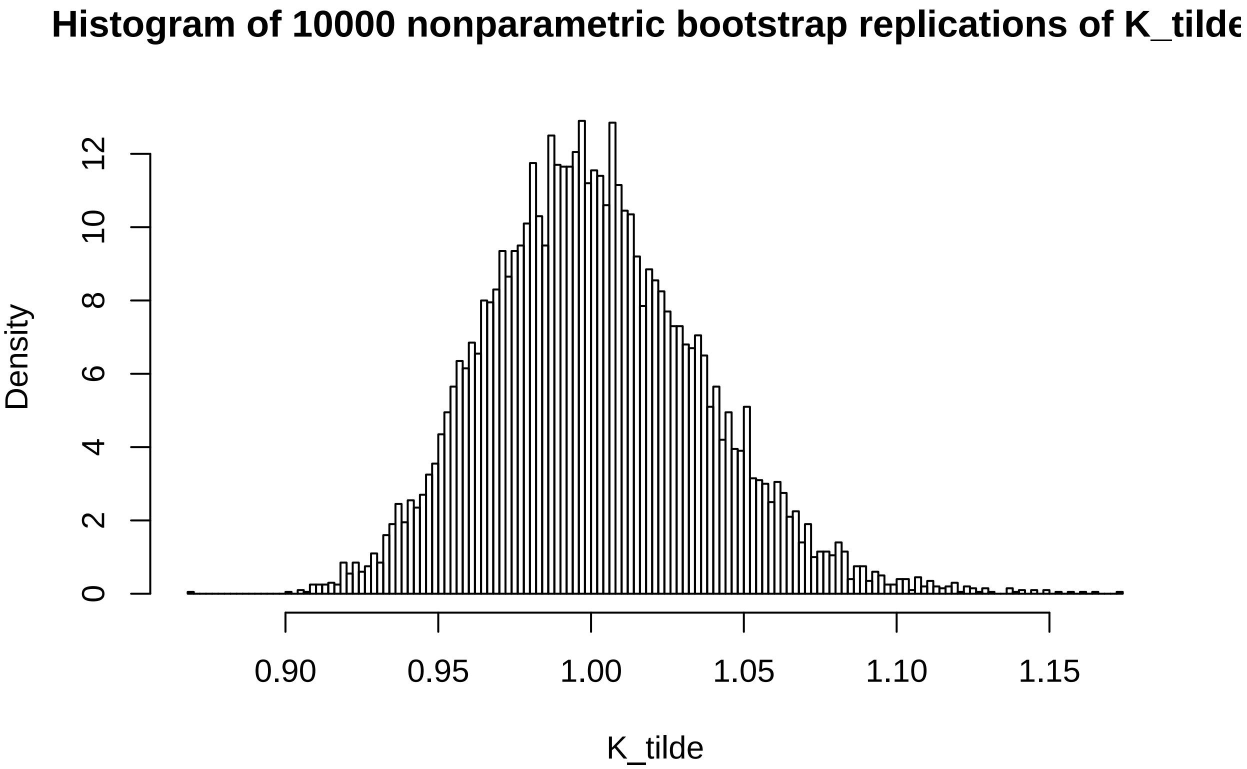
I will use the bootstrap method to formally test the null hypothesis above. In order to proceed with the sampling under H_0 I define $K = \frac{X}{U}$ and $\widetilde{K} = K - 0.239$. Under $H_0 : E(\widetilde{K}) = 1$ (using the first order approximation $E(K) = \frac{E(X)}{E(U)}$).

I want to estimate θ without assuming any parametric model for X and U. Thus, I perform non-parametric bootstrap,

```r
stat <- function(data, indices){

  d <- data[indices, ]
  k <- (mean(d$x)/mean(d$u)) - 0.239

  return(k)
}

set.seed(1235) # seed for the RNG to ensure that you get exactly the same results as here
boot_p <- boot(data=bigcity, statistic=stat, R=10000)
```

- first I define the statistic I want to compute in every bootstrap sample (i.e, `stat`). It is simply a function were I sample pairs of rows from my dataset and the compute $\widetilde{K}$.
- then I use the `boot()` function that actually performs the resampling; where I specify the dataset I want to sample from (i.e., `data=bigcity`), the statistic I want to compute in each sample (i.e.,`statistic=stat`) and how many bootstrap samples I want (i.e., `R=10000`)

```r
hist(boot_p$t, breaks=200, probability = T, xlab="K_tilde", main="Histogram of 10000 nonparametric boot
```



The figure shows the histogram of the sampled $\widetilde{K}$, under the null hypothesis. We see that the distribution is centered around 1, as we expected.

Next, I compute the Monte Carlo p-value as

$$p = \frac{1 + \#(|\theta^*| > |\widehat{\theta}|)}{R+1}$$

where θ^* is the bootstrap statistic.

```r
(p <- (1+sum(abs(boot_p$t) > abs(theta_hat)))/(10000+1))
```

```
## [1] 9.999e-05
```

Since $p < 0.005$ we reject the null hypothesis at 5% significance level and we conclude that there is a difference between the average population in the 49 cities between 1920 and 1930. Considering the higher number of people in 1930, we can deduce that the population has grown within the 10-year period. An informative approach would be to construct a CI for θ.

```r
#construct confidence intervals for the parameter estimate
meanratio <- function(data, indices){

  d <- data[indices, ]
  k <- (mean(d$x)/mean(d$u))

  return(k)
}

boot.out <- boot(bigcity, meanratio, 10000)
boot.ci(boot.out)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
```

```
##
## CALL :
## boot.ci(boot.out = boot.out)
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 1.167,  1.307 )   ( 1.161,  1.301 )
##
## Level     Percentile            BCa
## 95%   ( 1.177,  1.317 )   ( 1.179,  1.322 )
## Calculations and Intervals on Original Scale
```

- the code changes slightly because now I don't need to resample under the null hypothesis. Hence, $\widetilde{K}$ is just the ratio of the two means, $\bar{x}$ and $\bar{u}$.

We get 4 different kind of intervals. For their explanation I refer to Efron and Tibshirani (1994) and DiCiccio and Efron (1996).

All of them lead us to the same conclusion as before. Since the CI does not include one we can say that there is a difference between the two measurements. Moreover, considering that the CI is above one it allows us conclude that there has been a statistically significant growth in the population between 1920 and 1930.

As mentioned above the estimator we used is biased (Van Kempen and Van Vliet 2000). The bootstrap estimate of bias and SE are

```
boot_p
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = bigcity, statistic = stat, R = 10000)
##
##
## Bootstrap Statistics :
##     original      bias    std. error
## t1* 1.000019 0.001145971  0.03537191
```

The ratio of the estimated bias to standard error, $\widehat{bias}_{10000}/\tilde{SE}_{10000} = 0.03$ is small, indicating that in this case we do not have to worry about the bias. According to Efron and Tibshirani (1994) a bias of less than .25 standard errors can be ignored. That is because if bias/SE $\leq$ .25, then the root mean square error is no more than about 3.1% greater than SE.

Furthermore, to our convenience Van Kempen and Van Vliet (2000) provide a closed form expression for the bias of $\widehat{\theta}$

$$bias = \frac{1}{n}\Big(var(u)\frac{\mu_x}{\mu_u^3} - \frac{cov(x,u)}{\mu_u^2}\Big)$$

Using this formula we calculate $bias = 0.0016$. Notice $bias \simeq \widehat{bias}_{1000}$, an indication that our bootstrap algorithm approximated very close the theoretical result. The same is valid for the standard error: $\widehat{SE} = 0.034$ comparing to $\tilde{SE}_{10000} = 0.035$.

To conclude, the bootstrap approximated well the theoretical result and thus we can infer that there has been growth in the population of the US between 1920 and 1930.

Another way to test the same hypothesis

Under the null hypothesis (no growth) the mean of the difference is expected to be zero. Let $Z_i = X_i - U_i$, where i refers to the i^{th} city. Then

$$E(Z_i) = E(X_i - U_i) = 0.$$

Hence I define, $\widetilde{Z_i} = Z_i - E(Z) + E(Z_i)$

```
z <- bigcity$x-bigcity$u
mean(z)
```

```
## [1] 24.65306
```

```
(z.tilde <- z - mean(z))
```

```
##  [1] -19.6530612 -13.6530612 -16.6530612  56.3469388   2.3469388
##  [6]   1.3469388  -3.6530612   0.3469388  56.3469388  23.3469388
## [11] -10.6530612 -17.6530612 -16.6530612   7.3469388  -5.6530612
## [16]  -5.6530612 -16.6530612 -20.6530612  58.3469388  47.3469388
## [21]   3.3469388 -27.6530612 102.3469388 -10.6530612 -12.6530612
## [26] -11.6530612  -4.6530612 -21.6530612  23.3469388   7.3469388
## [31] -33.6530612 -14.6530612 -16.6530612 -24.6530612 -29.6530612
## [36] -13.6530612  -4.6530612  -5.6530612 -32.6530612 -10.6530612
## [41] -25.6530612  61.3469388  -0.6530612 -10.6530612  -6.6530612
## [46]  -6.6530612 -17.6530612  46.3469388  -6.6530612
```

```
mean(z.tilde)
```

```
## [1] -5.794596e-16
```

Now we can generate bootstrap sample under the null hypothesis

$$H_0 : \mu_Z = 0,$$
$$H_a : \mu_Z \neq 0$$

```
R = 10000
t.boot<-c(1:R)

for(b in 1:R){
  z.b <- sample(z.tilde, size=length(z), replace=TRUE)
  t.boot[b] <- t.test(z.b,mu=0)$statistic
}
```

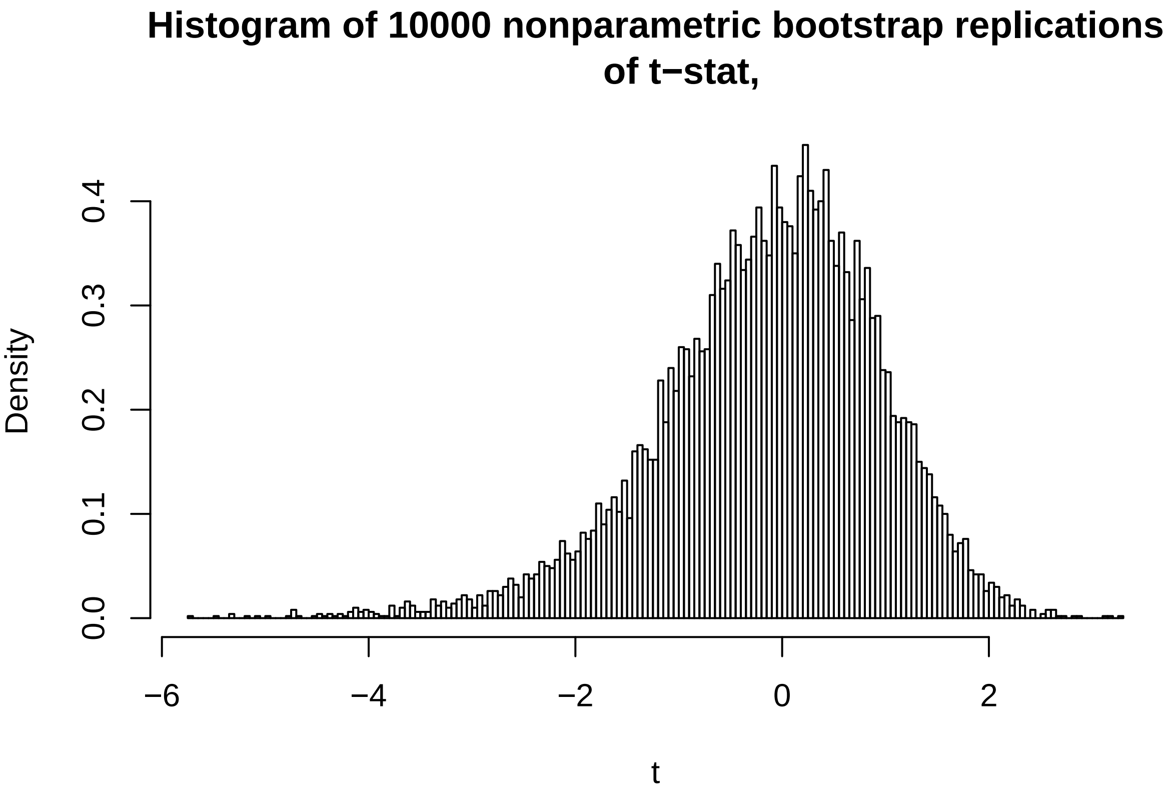
- I don't use the `boot()` function even though I could. I write my own for-loop.
- `z.b` samples 49 observations with replacement from `z.tilde` at each iteration (in total R=10000 iterations).
- `t.boot` calculates the following t-statistic at each iteration

$$t = \frac{\overline{z.b} - 0}{SE(\overline{z.b})},$$

where $\overline{z.b}$ is the mean of `z.b` at each iteration.

I plot the sampled t_s. We see that the distribution is centered around 0, as we expected.

```
hist(t.boot, breaks=200, probability = T, xlab="t", main="Histogram of 10000 nonparametric bootstrap rep
     of t-stat,")
```



The p-value is

```
(Pmc <- (1+sum(abs(t.boot) > abs(mean(z))))/(R+1))
```

```
## [1] 9.999e-05
```

As expected, it is exactly the same as before.

References

DiCiccio, Thomas J, and Bradley Efron. 1996. "Bootstrap Confidence Intervals." *Statistical Science*. JSTOR, 189–212.

Efron, Bradley, and Robert J Tibshirani. 1994. *An Introduction to the Bootstrap*. CRC press.

Van Kempen, GMP, and LJ Van Vliet. 2000. "Mean and Variance of Ratio Estimators Used in Fluorescence Ratio Imaging." *Cytometry* 39 (4). Wiley Online Library: 300–305.