Bootstrap Practical

STA Honours, Statistical Computing

University of Cape Town

last-modified 2024"

R: Some basic functions useful for bootstrapping

- Look up the function sample() in the R help.
- Use it to generate a random sample $\mathbf{u} = u_1, \ldots, u_{50}$ (of size 50) of **integers** from a uniform - discrete distribution U(0,99), call this u.
- Use sample() to draw a random sample of size 10 with and without replacement from u.
- Use sample() to draw 10 random samples of size 50 from u with replacement (like bootstrap sampling, also called resampling).

Prac 1: Median speed of galaxies

```
library(MASS)   # Modern Applied Statistics with S
                # (Venables and Ripley, 2004)

gal <- galaxies     # data set on velocities (km/s)
                    #of 82 galaxies

hist(galaxies, breaks = 20)
```

```
## MEDIAN speed of galaxies
tobs <- median(gal)       # T(X)
```

1. Set the number of bootstrap samples to $B = 10000$.
2. Take B bootstrap samples, each time calculate the median speed, store these in tboot.
3. What is `mean(tboot) - tobs`?
4. What is `sd(tboot)`? Will this decrease if B is increased? Explain. What exactly will this value tell us?

5. Plot the bootstrap distribution (histogram), indicate `tobs` on this.
6. Calculate a percentile and a basic bootstrap confidence interval for median galaxy speed. Compare.
7. Calculate percentile, basic bootstrap and bootstrap t confidence intervals for mean galaxy speed. Compare

Prac 2: galaxies again

```
library(boot)

bt.smpls <- boot(gal, function(x, i) median(x[i]), R = 3000)

# see Venables and Ripley, pg.134
# take 3000 bootstrap samples, returns the 3000 medians of these
```

Table 1: Relative Risk of Cardiovascular Disease

Blood Pressure	Cardiovascular Disease
High	$55/3338 = 0.0165$
Low	$21/2676 = 0.0078$
Relative risk	2.12

```
bt.smpls
summary(bt.smpls)

boot.ci(bt.smpls, type = c("norm", "basic", "perc", "bca"))     # all sorts of confidence intervals
help(boot.ci)            # bca gives a bias corrected and accelerated
                        # (improved) percentile interval
```

Compare with your values for bias, standard error, percentile and basic bootstrap CI

Prac 3: Regression Problem

For this question use the airquality data (`airquality` R data set).

For the correlation between ozone and temperature, find an estimate of SE, bias and construct a confidence interval. Use a nonparametric bootstrap.

Some R code that may be useful:

```
n <- dim(df)[1]

# sample cases from data frame
cases <- sample(1:n, replace = T)
booti <- df[cases, ]
```

Extra: Use parametric bootstrapping to construct a confidence interval for the correlation.

Prac 4: Relative Risk

Table 1 gives rates of cardiovascular disease for subjects with high or low blood pressure. The high-blood pressure group was 2.12 times as likely to develop the disease.

Find the following (your answers should correspond, approximately, to the values in brackets):

- bias (0.11)
- bootstrap SE (0.62)
- percentile bootstrap interval: (1.3, 3.7)
- basic bootstrap interval

Hint: Observations are binary. There are two groups.

Is there an increased risk of cardiovascular disease with high blood pressure?

Is the estimate of relative risk biased?

Which of the two confidence intervals is better?