Exercise 4

Bootstraping Practicals I

Olesia Galynskaia 12321492

2025

Fix the random seed for reproducibility

```
set.seed(12321492)
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE, fig.width = 6, fig.height = 4, dpi
```

Task 7

Making a sample

```
x_t7 <- c(82, 107, 93)
```

1. Mean of the sample

The sample contains the values 82, 107, and 93.
The mean of this sample is:

$$\bar{x} = \frac{82 + 107 + 93}{3} = 94.$$

```
mean_t7 <- mean(x_t7)
mean_t7
```

```
## [1] 94
```

2. Number of possible bootstrap samples + list all triplets

We take a bootstrap sample of size 3 with replacement.
Each position can be 82, 93, or 107.
This gives $3^3 = 27$ possible combinations.
`expand.grid()` prints all 27 triplets.

```r
# Generate all possible bootstrap samples (all triplets with replacement)
all_triplets <- expand.grid(x_t7, x_t7, x_t7)

# Count how many there are
n_triplets <- nrow(all_triplets)

n_triplets
```

```
## [1] 27
```

```r
all_triplets
```

```
##    Var1 Var2 Var3
## 1    82   82   82
## 2   107   82   82
## 3    93   82   82
## 4    82  107   82
## 5   107  107   82
## 6    93  107   82
## 7    82   93   82
## 8   107   93   82
## 9    93   93   82
## 10   82   82  107
## 11  107   82  107
## 12   93   82  107
## 13   82  107  107
## 14  107  107  107
## 15   93  107  107
## 16   82   93  107
## 17  107   93  107
## 18   93   93  107
## 19   82   82   93
## 20  107   82   93
## 21   93   82   93
## 22   82  107   93
## 23  107  107   93
## 24   93  107   93
## 25   82   93   93
## 26  107   93   93
## 27   93   93   93
```

3. Mean of each bootstrap sample

Here we take every bootstrap triplet and compute its mean.
`rowMeans()` gives the average of each row in the table, so we get 27 mean values — one for each possible bootstrap sample.

```r
bootstrap_means <- rowMeans(all_triplets)

bootstrap_means
```

```
##  [1]  82.00000  90.33333  85.66667  90.33333  98.66667  94.00000  85.66667
##  [8]  94.00000  89.33333  90.33333  98.66667  94.00000  98.66667 107.00000
## [15] 102.33333  94.00000 102.33333  97.66667  85.66667  94.00000  89.33333
## [22]  94.00000 102.33333  97.66667  89.33333  97.66667  93.00000
```

4. Mean of all bootstrap means

The average of all 27 bootstrap means is basically the same as the original sample mean.
Here both values are 94.
For the sample mean, the bootstrap does not introduce bias, so the two match.

```
orig_mean <- mean(x_t7)
mean_bootstrap_means <- mean(bootstrap_means)

orig_mean
```

```
## [1] 94
```

```
mean_bootstrap_means
```

```
## [1] 94
```

5. Fit a normal distribution to the sample

Here we pretend the data comes from a normal distribution.
To "fit" this normal distribution, we just take the sample mean and the sample standard deviation.
These two numbers fully define the normal model.

```
norm_mean <- mean(x_t7)
norm_sd   <- sd(x_t7)

norm_mean
```

```
## [1] 94
```

```
norm_sd
```

```
## [1] 12.52996
```

6. Draw 27 normal samples of size 3 and compute their means

We use the normal model fitted in step 5 (same mean and sd as the sample).
Then we generate 27 new samples of size 3 from this normal distribution.
For each sample we compute the mean.
These normal-based means should be close to the original mean (94) on average, just like the bootstrap means.

```r
n_samples <- 27

normal_means <- replicate(n_samples, mean(rnorm(3, mean = norm_mean, sd = norm_sd)))

normal_means
```

```
##  [1]  97.40850  74.21974  99.32222  83.42125  98.35519  84.71206  89.88526
##  [8]  98.45196  95.61571  84.39308  80.74189  90.32327  97.05686 100.21947
## [15]  94.13746 107.69425  98.29528  99.69141  92.44394  93.30317  95.36638
## [22]  75.38212  91.26943  94.80599  93.94257  92.41392  99.46622
```

7. Low and high values of the means

The bootstrap means go from 82 to 107, because these are the smallest and largest possible values we can get when resampling the original numbers.
The normal-based means have a wider range (in my run about 74 to 108), because the normal distribution is continuous and can produce values outside the original sample.

```r
# Range of bootstrap means (from all 27 triplets)
range_bootstrap <- range(bootstrap_means)

# Range of normal-based means (27 samples generated in step 6)
range_normal <- range(normal_means)

range_bootstrap
```

```
## [1]  82 107
```

```r
range_normal
```

```
## [1]  74.21974 107.69425
```

Task 8

Making a sample

```r
x_t8 <- c(4.94, 5.06, 4.53, 5.07, 4.99, 5.16,
          4.38, 4.43, 4.93, 4.72, 4.92, 4.96)

n_t8 <- length(x_t8)
n_t8
```

```
## [1] 12
```

1. Number of possible bootstrap samples

A bootstrap sample has size 12 and we sample with replacement from 12 values.
Each of the 12 positions can be any of the 12 data points, so the total number of possible bootstrap samples is 12^{12}.

```r
n_boot_t8 <- n_t8^n_t8
n_boot_t8
```

```
## [1] 8.9161e+12
```

2. Mean of each bootstrap sample

Here we just compute the usual sample mean of the 12 observations.
This value will be our reference when we look at the bootstrap means.

```r
mean_t8 <- mean(x_t8)
mean_t8
```

```
## [1] 4.840833
```

3. 2000 bootstrap samples and their means

We draw 2000 bootstrap samples from `x_t8`.
Each bootstrap sample has size 12 and is drawn with replacement.
For every resample we compute the mean, so `boot_means_t8` contains 2000 bootstrap means.

```r
B_t8 <- 2000

# 2000 bootstrap means: each time we resample 12 values with replacement
boot_means_t8 <- replicate(B_t8, mean(sample(x_t8, replace = TRUE)))

# Look at the first few bootstrap means
head(boot_means_t8)
```

```
## [1] 4.830833 4.901667 4.815000 4.885000 4.755833 4.879167
```

We do not create new bootstrap samples here.
We just average:
- the first 20 bootstrap means,
- the first 200 bootstrap means,
- and all 2000 bootstrap means.

As we use more bootstrap means, these averages get closer to the sample mean from step 2.

3.1 Mean of first 20

```r
mean_20   <- mean(boot_means_t8[1:20])

mean_20
```

```
## [1] 4.857917
```

3.2 Mean of first 200

```r
mean_200   <- mean(boot_means_t8[1:200])

mean_200
```

```
## [1] 4.83845
```

3.3 Mean of 2000

```r
mean_2000 <- mean(boot_means_t8)

mean_2000
```

```
## [1] 4.84126
```

4. Relate all the different bootstrap means to the sample mean

Here we put everything side by side:
the original sample mean and the three averages of bootstrap means.

The averages based on 20 and 200 bootstrap means are close to the sample mean but still a bit noisy.
The mean based on all 2000 bootstrap means is almost identical to the sample mean, which shows that the bootstrap distribution is centered around it.

```r
c(sample_mean = mean_t8,
  mean_20     = mean_20,
  mean_200    = mean_200,
  mean_2000   = mean_2000)
```

```
## sample_mean     mean_20    mean_200   mean_2000
##    4.840833    4.857917    4.838450    4.841260
```

5. Bootstrap 95% intervals from 20, 200 and 2000 means

```r
ci_20   <- quantile(boot_means_t8[1:20],   probs = c(0.025, 0.975))
ci_200  <- quantile(boot_means_t8[1:200],  probs = c(0.025, 0.975))
ci_2000 <- quantile(boot_means_t8,         probs = c(0.025, 0.975))

ci_20
```

```
##     2.5%    97.5%
## 4.754521 4.945958
```

```r
ci_200
```

```
##     2.5%    97.5%
## 4.706292 4.966667
```

```
ci_2000
```

```
##     2.5%    97.5%
## 4.695792 4.972521
```

We use the empirical 2.5% and 97.5% quantiles of the bootstrap means to build 95% intervals.

With only 20 bootstrap means the interval is quite rough and unstable.
With 200 bootstrap means the interval is more stable.
With 2000 bootstrap means the interval changes only slightly compared to 200; it looks much more reliable.

The intervals from 200 and 2000 means are very similar (roughly [4.70, 4.97]), which shows that 2000 bootstrap samples are enough for a stable estimate.

So more bootstrap samples give a smoother and more trustworthy interval.

6. t-test based 95% CI and comparison

```
t_ci_t8 <- t.test(x_t8, conf.level = 0.95)$conf.int
t_ci_t8
```

```
## [1] 4.674344 5.007323
## attr(,"conf.level")
## [1] 0.95
```

`t.test()` gives the usual 95% confidence interval for the mean based on a normal / t-distribution assumption. In this example t-based interval [4.67, 5.01] is slightly wider than the bootstrap intervals (about [4.71, 4.97] and [4.70, 4.97]).

The 95% bootstrap intervals from 200 and 2000 means are a bit shorter and sit in a similar region.
All intervals contain the sample mean and are close to each other,
so the bootstrap and the t-based method tell a consistent story about the mean.