IX1501 HT23 Project 2

Bootstraping for parameter estimation

Course code: IX1501
Date: 2023-10-05

Samir Alami, samirala@kth.se
Ali Sahibi, sahibi@kth.se

Summary

In this project we were given a sample stick of 10 observations. The observations are samples from 10 Random Variables(RV) which are independent and identically distributed. We do not know the probability distribution of the RV’s and the mean μ is unknown as well.

Our first task is to describe how the bootstraping method can be used to estimate the probability:

$$p = P\left(-5 < \frac{\sum_{i=1}^{n} X_i}{n} - \mu < 5\right)$$

And then we estimate p using the described method. And p was estimated to:

```
In[695]:=
Dynamic[p]
Out[695]=
0.761464
```

Method

First we estimate μ by taking the mean of the sample that we have. After estimating μ we can simply the expression

$$p = P\left(-5 < \frac{\sum_{i=1}^{n} X_i}{n} - \mu < 5\right) = P\left(-5 < \overline{X} - \mu < 5\right)$$

to

$$p = P\left(\mu - 5 < \overline{x} < \mu + 5\right) = P\left(71.7 < \overline{x} < 81.7\right)$$

Now we can use bootstrapping to estimate p. Through bootstrapping we can estimate the distribution of the mean $\overline{x}$.

Here’s pseudocode description of the bootstrapping method:

```
m = number of resamples to generate.
samplelist = list of sampled means. Initially empty.

for m loops do:
        resample = create a random resample of 10 elements using existing stick sample.
        mean = calculate the mean value of the sample.
        samplelist->Add(mean) = add the mean to the list of sampled means.

sorted_samplelist = Sort the list of sampled means.
Return sorted_samplelist.
```

Now we have a large set of ordered means from random resamples. This give us a distribution of the mean value $\overline{x}$. To estimate p we need to create a CDF function.

```
cdf(x) = Sum(elements in sorted_samplelist <= x)/Sum(all elements in sorted_samplelist)
```

Now we can estimate p as:

```
p = cdf(μ+5) - cdf(μ-5)
```

Result

```
In[696]:=
 Dynamic[p]
Dynamic[percentp]
Out[696]=
0.761464
Out[697]=
76.15%
```

Above value is the estimation of p using bootstraping.

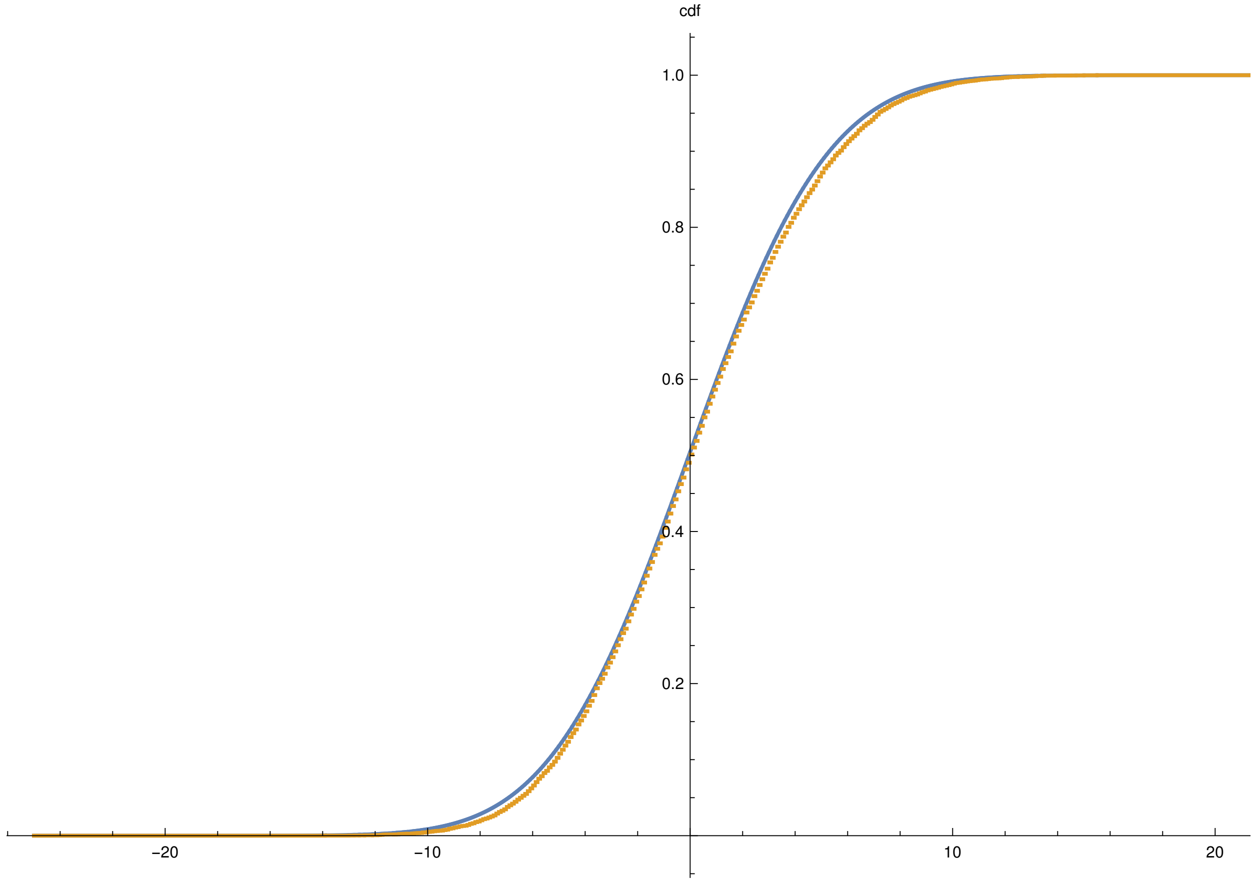
Discussion

Essentially what p represents is the degree of confidence for the confidence interval CI $= (\overline{x} - \mu) \pm 5$, where $\overline{x} - \mu$ represents the difference between the true mean $\overline{x}$ and the stick sample estimation μ,p is the degree of confidence that the absolute error between the true mean and estimated mean is < 5. After estimating the sample mean μ, 76.7, we can rewrite the probability to $P(\mu - 5 < \overline{x} < \mu + 5) = P(71.7 < \overline{x} < 81.7)$. This expression will give us the same probability and be easier to deal with when bootstrapping.

The resulting value we got for p is `0.761464`. So the absolute Error between the true mean and the estimated mean μ ,76.7, based on the stick sample is < 5 with a confidence of `76.15%`.

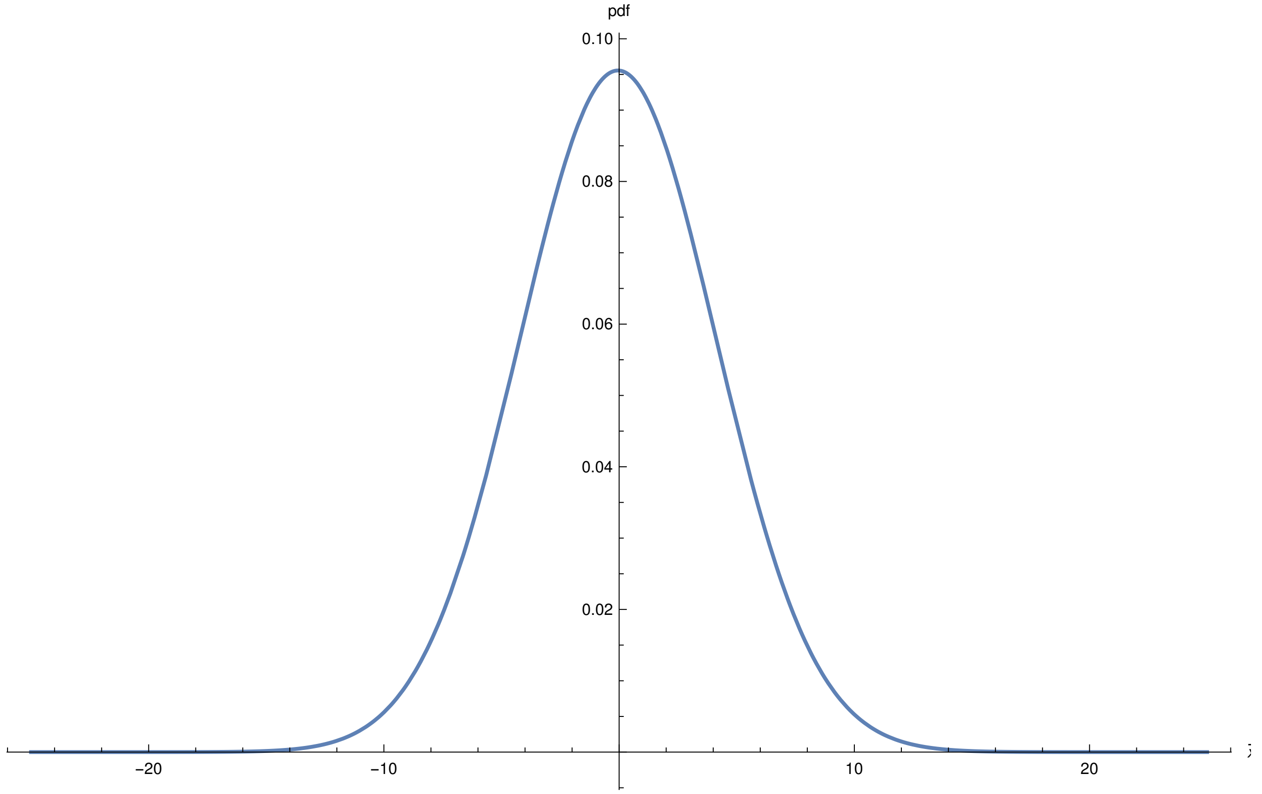
With the bootstraping method we get the mean $\overline{x}$ distribution. Since we know that each resampled sample from bootstrapping consists of random and independent observations of RV's, the mean distribution of the samples is normally distributed according to CLT. Therefore our set of means from bootstrapping must be normally distributed, and by taking the sample mean and sample standard deviation of the means we can get the normally distributed pdf and cdf of the mean distribution. Below is a comparison of the normally distributed cdf and the cdf function based on the set of means:

Out[∘]=



The functions match up well. And using the normally distributed cdf function we get an estimation off p as `0.768893`. We can also plot the probability distribution function of the absolute error between the true mean and estimated mean:

Out[ ]=



Code

The sample.

In[698]:=

```
sample = {56, 101, 78, 67, 93, 87, 64, 72, 80, 69};
```

Estimating μ:

In[699]:=

```
μ = N[Mean[sample]]
```

Out[699]=

```
76.7
```

we Realize p=$P\left(-5 < \overline{x} - \mu < 5\right) = P(-5 + \mu < \overline{x} < 5 + \mu)$

In[700]:=

```
low = -5 + μ
top = 5 + μ
```

Out[700]=

```
71.7
```

Out[701]=

```
81.7
```

we further realize that $p = P(\text{low} < \overline{x} < \text{top})$.

Below is a bootstraping function which takes a sample and generates a number of resamples and saves the mean for each resample in a list. The function then returns the sorted list of mean values.

```
In[702]:=
BootStrap[sample_, Nresamples_] := Module[{n = Nresamples, means = {}},
       For[i = 1, i ≤ n, i++,
            randResample = RandomChoice[sample, Length[sample]];
            means = Join[means, {N[Mean[randResample]]}];
        ];
       Return[Sort[means]];
  ];
```

A cdf function of the bootstrap mean distribution.

```
In[703]:=
ecdf[x_, means_] := N[Sum[If[i ≤ x, i, 0], {i, means}] / Sum[i, {i, means}]];
```

A cdf estimation function based on normaldistribution.

```
In[704]:=
pdf[x_, means_] := N[PDF[NormalDistribution[Mean[means], StandardDeviation[means]], x]];
In[705]:=
cdf[x_, means_] := N[CDF[NormalDistribution[Mean[means], StandardDeviation[means]], x]];
```

We generate a set of means.

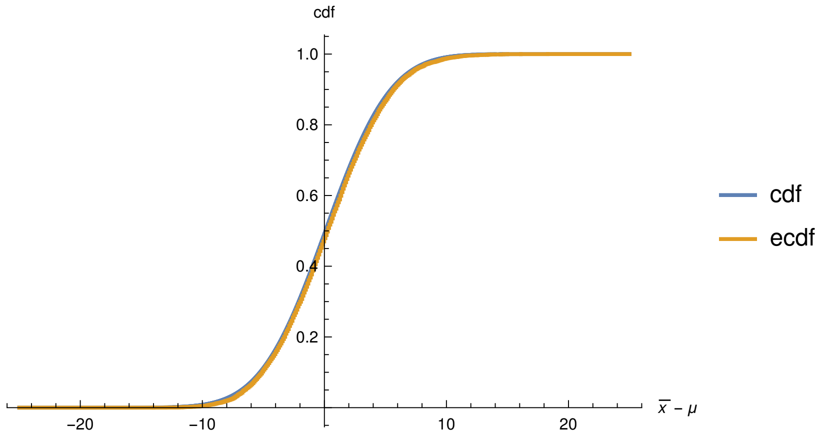
```
In[706]:=
nresamples = 10 000;
In[707]:=
means = BootStrap[sample, nresamples];
```

We are interested in estimating $p = P\left(\text{low} < \overline{x} < \text{top}\right) \approx \text{ecdf(top)} - \text{ecdf(low)}$

```
In[708]:=
p = ecdf[top, means] - ecdf[low, means]
pN = cdf[top, means] - cdf[low, means]
Out[708]=
0.761464
Out[709]=
0.767091
In[710]:=
percentp = Dynamic[PercentForm[p]]
Out[710]=
76.15%
```

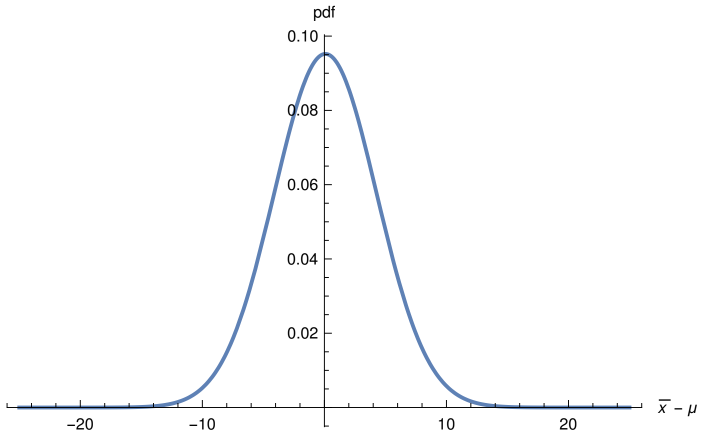
We plot the cdf based on the set of means(ecdf) as well as the normal distribution CDF and PDF based on the set of means parameters.

```
In[711]:=
plot0 = Plot[{cdf[x + μ, means], ecdf[x + μ, means]},
  {x, -25, 25}, AxesLabel → {"x̄ - μ", "cdf"}, PlotLegends → {"cdf", "ecdf"}]
Out[711]=
```



Plotting the pdf of the error rate distribution.

```
In[712]:=
plot1 = Plot[pdf[x + μ, means], {x, -25, 25}, AxesLabel → {"x̄ - μ", "pdf"}]
Out[712]=
```


pdf
$\bar{x} - \mu$