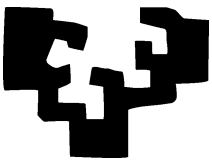
eman ta zabal zazu



Universidad del País Vasco

Euskal Herriko Unibertsitatea

Seminar 2

1 Synopsis.

What we are set about to do. In a previous seminar you saw a simple example on how to estimate things by simulation, using the Monte Carlo method. We will again resort to the same technique to study empirically the properties of different estimators.

We already have some theory and results concerning properties of estimators. However, in many cases, theory is intractable or only can shed light on what happens with large sample sizes. Simulation, therefore, is of great help to ascertain what happens with small samples, when the theory is too involved or simply not at all available.

What you need. You need to be fully acquainted with the content of the previous seminar and practice assignment. You will also need access to a computer equipped with R.

2 Background

We have seen theoretically that in a large number of cases, moment estimators and maximum likelihood estimators of the mean of a distribution are coincident. Although we also have seen one exception (the best estimator of the $\theta/2$, the mean of a uniform $U(0,\theta)$, is *not* $\overline{X} = n^{-1}(X_1 + \ldots + X_n)$), you might be tempted to think that for all practical purposes $\overline{X}$ is just "the right" estimator of the mean of a distribution.

In this seminar we will see that, in some cases, what would seem the obvious estimator of a location parameter has dismal performance; it may even fail to be consistent.

3 An example of total failure of $\overline{X}$ as estimator

3.1 The Cauchy distribution

We have met the Cauchy distribution in class. It has density,

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}; \tag{1}$$

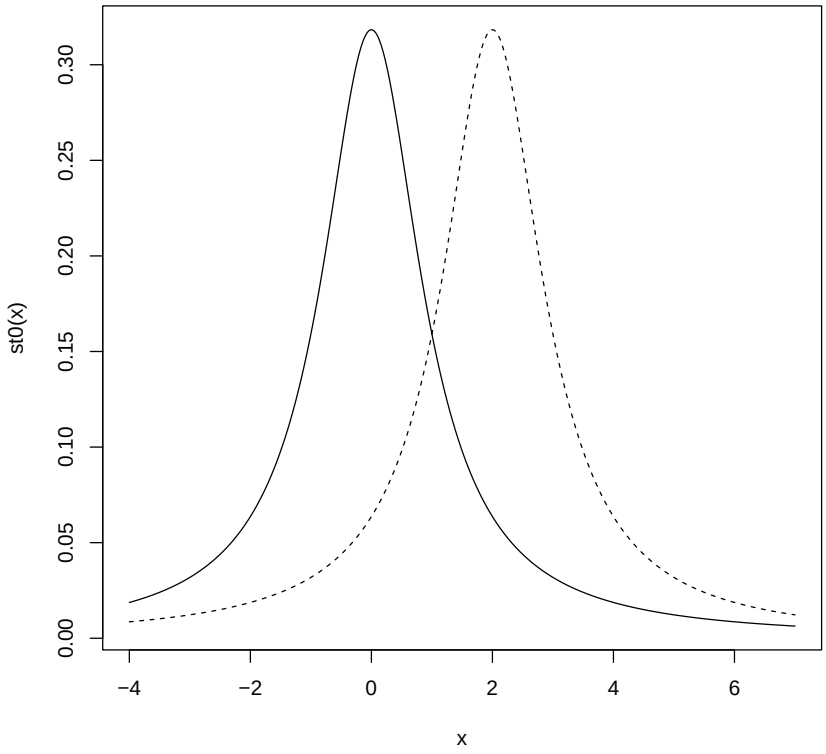
it is, with another name, the one-degree-of-freedom Student's t.

Consider now the shifted Cauchy distribution, with location parameter d. Its density is:

$$f(x) = \frac{1}{\pi}\frac{1}{1+(x-d)^2}; \tag{2}$$

it is the same distribution, translated d units to the right. See in Figure 1 the ordinary and shifted distributions when $d = 2$; d is called the *location parameter* of the distribution.

Figure 1: Ordinary and shifted Cauchy distributions (= Student's t_1 distributions). The dashed density has location parameter $d = 2$.



3.2 Computation of the likelihood

The likelihood for a sample of $x_1, \ldots, x_n$ is quite easy to write:

$$\ell(d; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\pi}\frac{1}{1+(x_i-d)^2} \tag{3}$$

Since the density is symmetric around the location parameter d, we might hope that $\overline{X} = (X_1 + \ldots + X_n)/n$ would be a "good" estimator of d, perhaps the MLE. Let's check that this is not the case.

We can easily generate a random sample of size $n = 2$ from the density (2), since it is just a t_1 shifted 2 units to the right:

```
> x <- rt(n=2,df=1) + 2
```

Each time we run the code above, we obtain two different observations. In this case,

```
> x

[1]  1.190118 -2.955073
```

Next we have to compute the likelihood (3) associated to that sample. Since this will have to be evaluated time and again, we can define a function as follows:

```
> lik <- function(d) { 1 / (  (1 + (x[1]-d)^2)  * (1 + (x[2]-d)^2) * pi^2 )  }
```

Note that `x[1]` and `x[2]` are the two values of the sample just generated. Each time we invoke function `lik` with an argument `d` we compute the likelihood of such `d` *given the sample*.

```
> lik(3)

[1] 0.0006498971
```

Observe that we could also compute the likelihood taking advantage of the pre-defined density function for the Student's t_1, which is just another name for the Cauchy distribution. Lets define

```
> lik.alternative <- function(d) { return(dt(x[1]-d,df=1) * dt(x[2]-d,df=1)) }
```

We can check that we obtain the same value before for the likelihood:

```
> lik.alternative(3)

[1] 0.0006498971
```

The MLE associated to the given sample is the value of d for which the function `lik` (or the equivalent `lik.alternative`) attains its maximum. Even with only two observations, it is messy to solve for that maximum. It will be much easier to plot the function over a range likely to contain the maximum and locate it visually. We can do something like:

```
> curve(lik,from=-5,to=5,n=1000)
> abline(v=mean(x))
> abline(v=2,col="red")
> text(2.15, 0, "d", col="red")
> text(mean(x)+0.2, 0, expression(bar(X)))
```

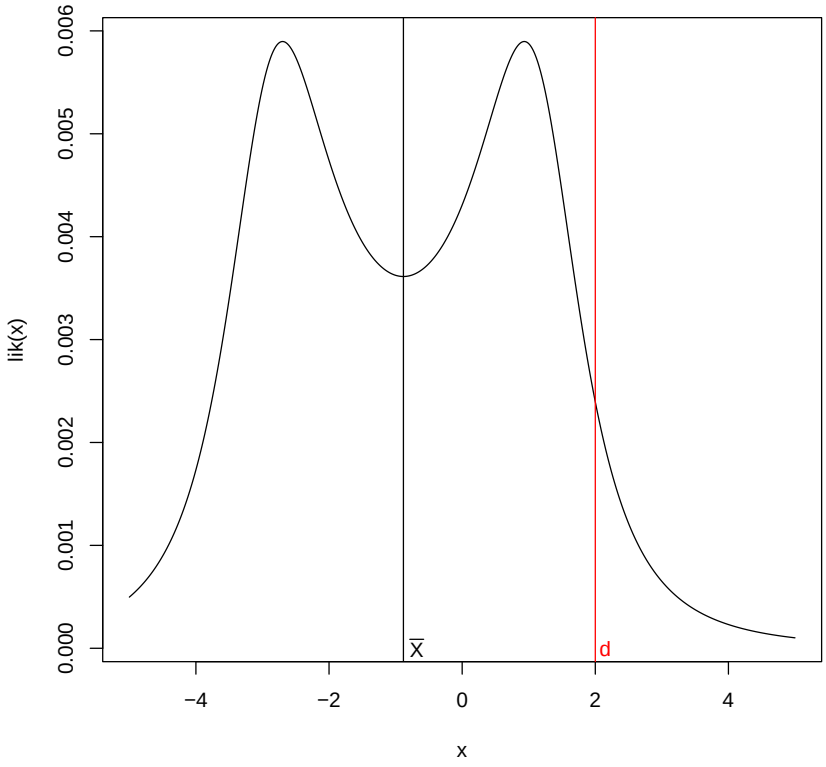
The result can be seen in Figure 2. In this case, we have two equal maxima —a not infrequent occurrence with this distribution, even with larger samples. Notice that $\overline{X} \neq \hat{d}_{\text{MLE}}$.

3.3 Properties of $\overline{X}$ as estimator of d

We may now turn to investigate the properties of $\overline{X}$ as estimator of d. First and foremost, is it consistent? It can be shown analytically that it is not; we will see what happens empirically as we let the sample size grow. First, let's generate a large number of observations from a shifted Cauchy:

```
> N <- 10000
> obs <- rt(n=N,df=1) + 2
```

Figure 2: Likelihood associated to the sample `x`. The red line marks the true d, the black line marks $\overline{X}$.



Next, we will compute $\overline{X}$ for samples of increasing size, and store the results in vector `Xn`, which we define beforehand and fill with zeroes:

```
> Xn <- rep(0,N)
> for (i in 1:N) {
    Xn[i] <- sum(obs[1:i]) / i
    }
```

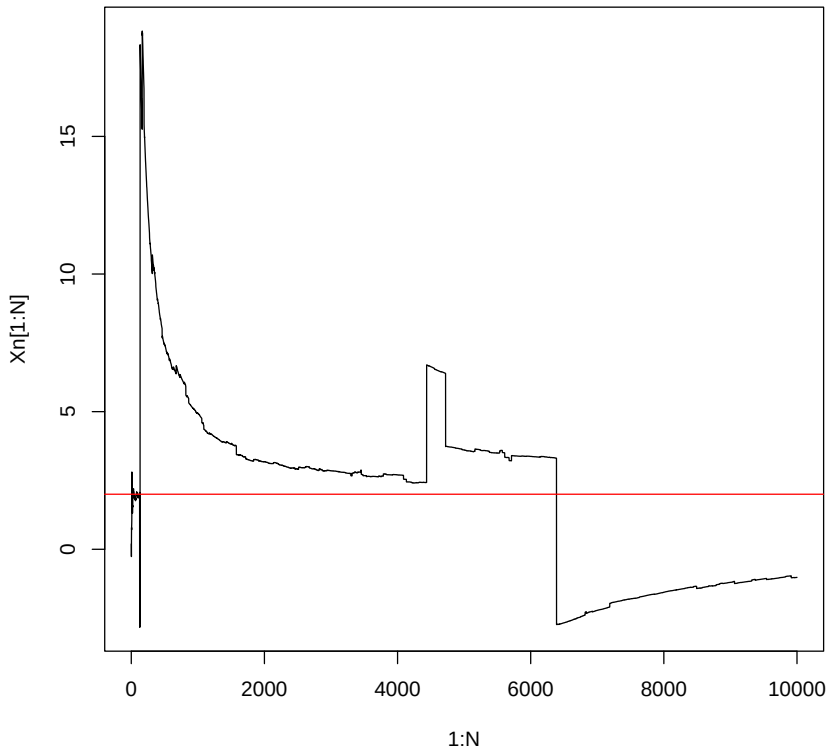
After executing this code[1], `Xn[n]` contains $\overline{X}_n$, the average of the first n observations. We can plot $\overline{X}_n$ versus n to see whether it approaches d:

```
> d <- 2
> plot(1:N,Xn[1:N],type="l")
> abline(h=d,col="red")
```

The result is shown in Figure 3. We see that there are no signs of $\overline{X}_n$ approaching d as n grows. In fact, $|\overline{X}_n - d|$ is larger for $n = 10000$ than if was for $n = 4000$. Looking at the large jumps of $\overline{X}_n$ we may suspect what the problem is: the Cauchy has thick tails, and every once in a while we get a

[1] Writen in the form above for clarity; it would be much faster to write `Xn <- cumsum(obs) / (1:N)`.

Figure 3: $\overline{X}_n$ does not appear to consistently estimate d, whose value is represented as a red line



very large observation which drastically changes the value of $\overline{X}_n$. In fact, $\overline{X}_n$ has no mean and its variance does not decrease with n[2].

3.4 What alternatives do we have?

A simple estimator of the location d like $\overline{X}_n$ completely fails. How then can we estimate d?

3.4.1 The maximum likelihood estimator

The maximum likelihood estimator enjoys in this case the usual good (large sample) properties, but it is complicated to obtain for all but small samples. The log likelihood is,

$$L(d; x_1, \ldots, x_n) = \log\left(\prod_{i=1}^{n} \frac{1}{\pi} \frac{1}{1+(x_i-d)^2}\right) \quad (4)$$

$$= -\sum_{i=1}^{n} \log(1+(x_i-d)^2) - n\log(\pi) \quad (5)$$

[2]It can be shown that the distribution of $\overline{X}_n$, for whichever value of n, is Cauchy (the same as the distribution of a single observation!), hence with no mean and infinite variance.

Taking the derivative with respect to d and equating to zero yields

$$\sum_{i=1}^{n} \frac{x_i - d}{1 + (x_i - d)^2} = 0 \tag{6}$$

which usually has to be solved numerically. The Cramér-Rao lower bound can be computed analytically[3]:

$$-\frac{\partial^2 \log f(x;d)}{\partial d^2} = \frac{2(1-(x-d)^2)}{(1+(x-d)^2)^2} \tag{7}$$

$$I(d) = E\left(\frac{2(1-(x-d)^2)}{(1+(x-d)^2)^2}\right) \tag{8}$$

$$= \int_{-\infty}^{\infty} \frac{1}{\pi} \frac{1}{1+(x-d)^2} \frac{2(1-(x-d)^2)}{(1+(x-d)^2)^2} dx \tag{9}$$

$$= \frac{1}{2}, \tag{10}$$

hence the Cramér-Rao lower bound for an estimator based in n observations is $2/n$ and the MLE converges in distribution to $N(d, \sigma^2 = 2/n)$ as $n \to \infty$.

3.4.2 Trimmed means

Since the problem with the instability of $\overline{X}_n$ as estimator of d seems to be the appearance of very large observations from time to time, we might think of computing the average dropping those observations. For instance, we could drop the 10% largest and smallest observations and compute the average of the remaining 80%. The resulting estimator is a so-called *trimmed mean*, with 10% trimming on each side, usually denoted as $\overline{X}_{[0.10]}$. It can be computed in R as,

```
> mean(x,trim=0.10)
```

As the trimming proportion approaches 50%, the trimmed mean approaches a median. Unlike the ordinary average $\overline{X}_n$ the median (and trimmed means) have finite variances and are consistent estimators of d.

3.5 Efficiency

Theoretical results are available on the variance of different trimmed means, so their efficiency is known[4]. In keeping with our empirical, experimental approach, we will approximate the efficiencies of $\overline{X}_{[0.10]}$ and the median by computing their mean square error for different values of n and comparing with the Cramér-Rao lower bound. Both are unbiased for large n.

We will consider sample sizes going from

```
> first <- 200
```

to

```
> N <- 2000
```

[3] The integration necessary for the expectation in (10) is not trivial.

[4] See for instance http://www.johndcook.com/Cauchy_estimation.html.

First, we define a matrix in whose N rows and two columns we will store the sample variances of the chosen estimators:

```
> VAR <- matrix(0,N,2)
```

Next, for each sample size $n = 200, \ldots, N$ we will generate $K = 60$ samples on size n. For each of the $K = 60$ samples, we compute the value of both estimators and store both values $(\hat{d} - d)^2$ in one row of `SE`. The average of the $K = 60$ square errors goes into the n-th row of `VAR`; this will be our approximation of the variance of each estimator for samples of size n.

```
> K     <-   60
> SE   <- matrix(0,K,2)
> for (n in first:N) {
    for (k in 1:K) {
      obs <- rt(n=n,df=1) + 2
      SE[k,1] <- (mean(obs,trim=0.10) - d)^2
      SE[k,2] <- (median(obs) - d)^2
    }
    VAR[n,] <- colMeans(SE)
  }
```

We can now plot the results:

```
> plot(first:N,VAR[first:N,1], col="black", type="l",
       xlab="Sample size", ylab="Estimated variance")
> lines(first:N,VAR[first:N,2],col="green")
```

and add the asymptotic variance of the MLE estimator for comparison purposes:

```
> lines(first:N, 2 / first:N, col="red",lwd=2)
> legend("topright",
         legend=c("Trimmed 10%","Median","Cramer-Rao bound"),
         text.col=c("black","green","red"))
```

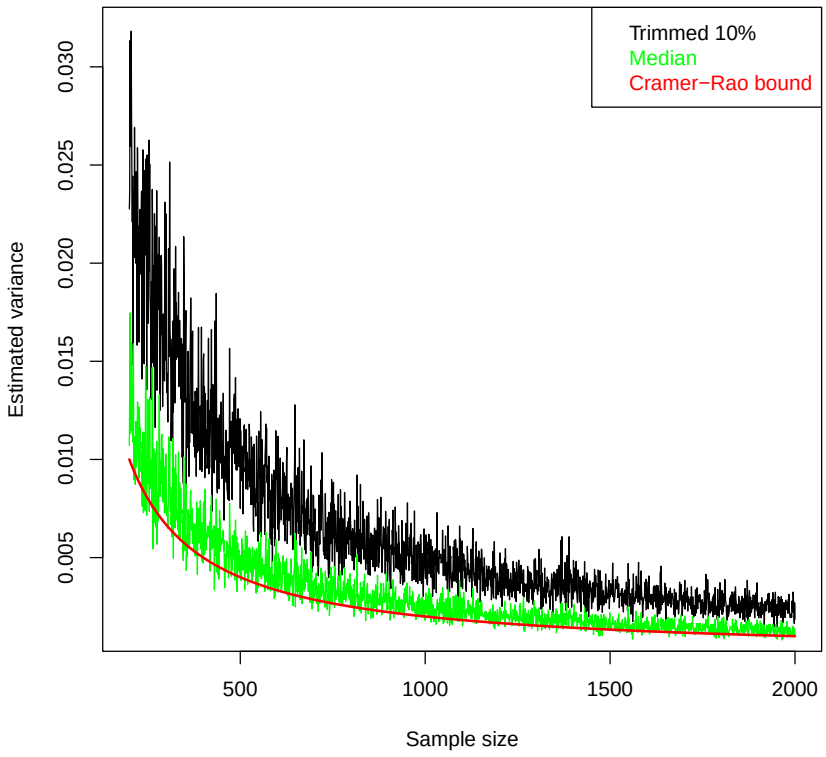
The results can be seen in Figure 4. We see that for a given sample size, $\overline{X}_{[0.10]}$ has larger variance than the median, and both are above the Cramér-Rao bound[5], so the median is more efficient in this case. An approximation to the efficiencies of both can be obtained dividing the CR bound of $2/n$ by the variance of either estimator for the same n. So for $n = 2000$, for instance,

```
> n <- 2000
> EffTrimmedMean <- (2/n) /  VAR[n,1]
> EffTrimmedMean

[1] 0.5391658

> EffMedian      <- (2/n) /  VAR[n,2]
> EffMedian

[1] 0.9705007
```

We see that the 10% trimmed mean has low efficiency, while the median reaches 97% of the theoretical optimum efficiency.

[5] Remember, though, that the variances displayed are estimated in a simulation, not real variances; it can happen that on occasion one of those estimated variances falls below the Cramér-Rao bound.

Figure 4: Estimation variance of $\overline{X}_{[0.10]}$ and the median as estimators of the location parameter d of a Cauchy. The red line gives the CR lower bound for each sample size.



4 Some further comments

1. The replacement of the mean by the median or something else is common in practice, whenever we deal with distributions with a very large variance. For instance, in sailing competitions ships usually race several times. For each ship, average points are computed dropping the best and worse races. The intent is to remove extreme observations which may be the outcome of very good (or bad) luck and keep the central ones, which presumably are a better indicator of the ability of crew.

2. Why use a trimmed mean in the case of sailing races and not the median? Did not our simulation show that the median is more efficient? That is the case for the Cauchy: but in other cases (e.g., the normal distribution) the ordinary mean is the most efficient. For situations half way between them, a trimmed mean may be a good compromise.

3. For an accessible account to some of the topics discussed you may turn to the Wikipedia, https://en.wikipedia.org/wiki/Cauchy_distribution