Computational Statistics

Lesson 7: On “Averaging”

Raymond Bisdorff

University of Luxembourg

December 6, 2019

Content of lecture 7

1. The benefit from averaging
   - The Law of Large Numbers
   - Estimate distribution parameters
   - Howto reduce noise – Graphical illustration
2. Convergence of the averaging
   - Convergence of the mean for a standard Gaussian
   - And if there are potential outliers?
   - Non convergence of the average for a Cauchy
3. Comparing two empiric means
   - Robustness of the *t* statistic
   - Estimate the t Statistics
   - Monte Carlo simulation of the H_0 rejection

The Law of Large Numbers

- Assume that $X_1, X_2, ..., X_n$ are independent and identically distributed random variables with finite mean μ and variance σ^2. Then

  $$\boxed{\lim_{n\to\infty} \frac{1}{n}\left(X_1 + X_2 + ... + X_n\right) \;=\; \mu}$$

  almost certainly.
- The expression *almost certainly* means that, with probability one, the averages of any realization $x_1, x_2, ...$ of the random varaibles $X_1, X_2, ...$ converge toward their common mean μ.
- This is good news, since many observed data sets concern multiple realizations of some random variables.

Estimating noise distribution parameters

- Assume that we have a measurement device whose output is a noisy signal; meaning that the signal observed contains a noise component.
- By attaching the device to a dummy load whose theoretical noiseless output we know, we may calibrate the proper noise level of the device, by subtracting this theoretical output, to obtain its pure noise level.
- We assume that an observed pure noise vector $x_1, x_2, ..., x_n$ contains n realizations of a same random variable $X : \Omega \Rightarrow \mathbb{R}$ with mean μ and variance σ^2.
- If we assume that the n realizations are mutually independent, the mean μ and variance σ^2 of X can be estimated via the formulas:

$$\hat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i\,, \quad \hat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2.$$

Howto reduce noise by averaging

- Having performed the noise calibration, we realize that the signal-to-noise ratio of our device is so poor that important features of the signal are cluttered under the noise.
- We may repeat the measurement *nSim* times and average the noisy signals in the hope of reducing the variance of the noise component.
- How many measurements do we need to reach a desired signal-to-noise ratio?

Howto reduce noise – continue

- If $x^k \in \mathbb{R}^n$ denotes the noise vector observed in the k-th repeated measurement, the average noise vector $\overline{x}$ becomes:

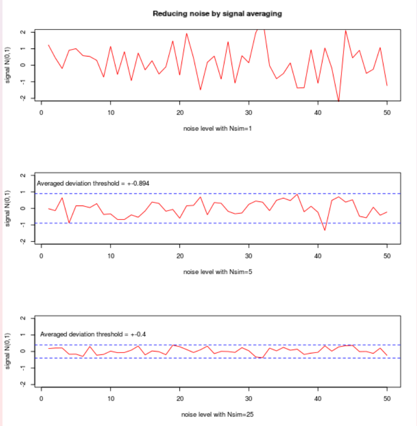
$$\overline{x} \;=\; \frac{1}{nSim} \sum_{k=1}^{nSim} \left(x^k\right) \;\in\; \mathbb{R}^n\,.$$

- From the CLT we know that, as $nSim \to \infty$, $\overline{x}$ becomes Gaussian distributed with mean μ and variance going to zero like $\sigma^2/nSim$.
- To obtain, hence, a signal whose variance is below a given threshold value τ^2, we need to choose $nSim$ such that: $\sigma^2/nSim \;<\; \tau^2$.

Howto reduce noise – Graphical illustration

To demonstrate the noise reduction by averaging, we generate 25 standard Gaussian noise vectors of dimension $n = 50$, and average them.

```
nSim=25
xn = rep(0,50)
for (i in 1:nSim) {
   xn = xn + rnorm(50,0,1)
}
xn = xn/nSim
plot(xn,type="l",ylim=c(-2,2),
  xlab="noise level with nSim=25",
  ylab="signal N(0,1)",col="red")
n = 2/sqrt(nSim)
abline(h=+n,lty=2,col="blue")
abline(h=-n,lty=2,col="blue")
```



Exercise

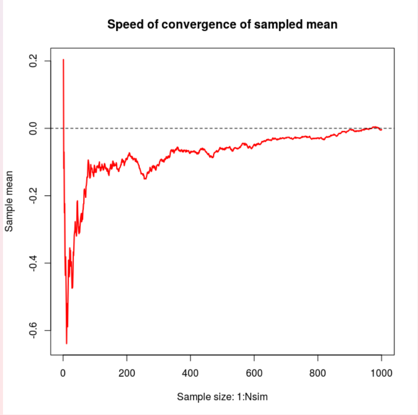
1. *Reconsider the previous noise reduction problem when observing a Gaussian noise X with estimated mean $\hat{\mu}$ and variance $\hat{\sigma}^2$. How many measurements nSim must be made in order to assure that* 95.5% *of the realizations will appear between $\mu \pm \hat{\sigma}$.*
2. *Realize a grahical illustration of your solution when assuming a standard Gaussian noise.*

1. The benefit from averaging
   - The Law of Large Numbers
   - Estimate distribution parameters
   - Howto reduce noise – Graphical illustration
2. Convergence of the averaging
   - Convergence of the mean for a standard Gaussian
   - And if there are potential outliers?
   - Non convergence of the average for a Cauchy
3. Comparing two empiric means
   - Robustness of the t statistic
   - Estimate the t Statistics
   - Monte Carlo simulation of the H_0 rejection

How fast converges the average?

Let us sample the mean of i.i.d. Gaussian variables $X_i \sim \mathcal{N}(\mu = 0, \sigma = 1)$. The LLN tells us that the sampled mean will approach certainly the common mean value 0 with a standard deviation $\sigma/\sqrt{nSim}$. How fast is this convergence ?

```
> nSim = 1000
> mn = rep(0,nSim)
> dn = rnorm(nSim)
> for (i in 1:nSim) {
+ mn[i] = mean(dn[1:i])}
> plot(mn,type="l",col="red")
> abline(h=0,lty=2)
```



And if there are potential outliers?

Let us now consider the ratio X/Y of two independent standard Gaussian variables $\mathcal{N}(\mu = 0, \sigma = 1)$. This ratio has a cumulative density function:

$$P[X/Y \leqslant z] = \frac{1}{2\pi}\int\int_{x/y\leqslant z} exp(-\frac{1}{2}x^2)\,exp(-\frac{1}{2}y^2)\,dxdy, \quad z \in \mathbb{R}.$$

which becomes a Cauchy density with mode 0 and spread 1:

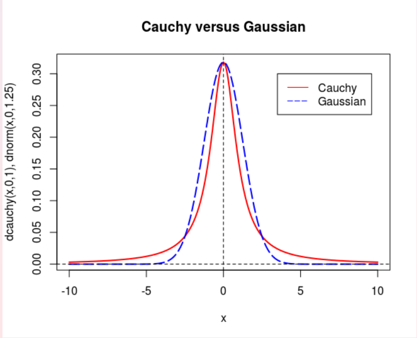
$$P[X/Y \leqslant z]'_z \;=\; \frac{1}{\pi(1+z^2)}, \quad z \in \mathbb{R}.$$

The Cauchy distribution, also called after Lorentz, has no finite mean and variance and the LLN will not work in this case.

The "*heavily tailed*" Cauchy distribution

The parameters of the Cauchy are the mode (called location in R) and a scale factor which may be aligned to match with the standard deviation concept. The solid red curve below is a Cauchy density with mode = 0 and scale 1. The dashed blue curve is a Gaussian density with same density peak at value $1/\pi$ at mean (or mode) 0 and standard deviation $\sqrt{\pi/2} = 1.253314$.

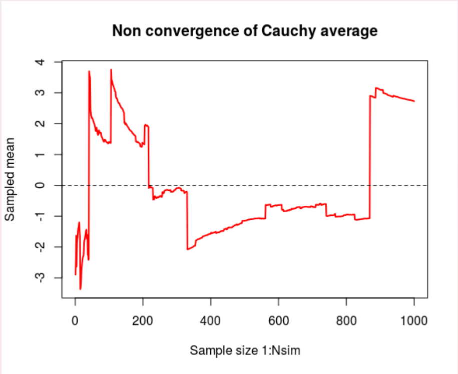
```
> x = seq(-10,10,by=0.1)
> plot(x,dcauchy(x,0,1),
+ "l",lwd=2,col="red")
> lines(x,dnorm(x,0,1.253),
+ lwd=2,lty=5,col="blue")
> abline(h=0,v=0,lty=2)
```



Non convergence of the Cauchy average?

Let us sample the mean of 1000 i.i.d. Gauchy variables with location = 0 and scale = 1.

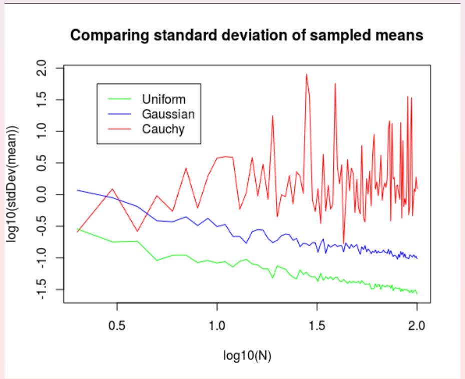
```
> nSim = 1000
> mn = rep(0,nSim)
> dc = rcauchy(nSim,0,1)
> for (i in 1:nSim) {
+ mn[i] = mean(dn[1:i])}
> plot(mn,type="l",
+ col="red")
> abline(h=0,lty=2)
```



Comparing sampled averages' standard deviations

The Figure below compares the standard deviation from samples of size $N = 1 : 100$ drawn from a standard uniform, Gaussian, and Cauchy random sequence of size 10 000. This is a log-log plot where $\sigma/\sqrt{N}$ appears as a more or less straight line with slope $\approx -1/2$ for the uniform and Gaussian distributions, wheras the Cauchy sample mean will evolve erratically.

```
> nSim = 10000
> sdc = rep(0,100)
> dc = rcauchy(nSim,0,1)
> for (N in 1:100) {
+  sampc = sample(dc,N)
+  sdc[N] = sd(sampc
+  - mean(sampc))/sqrt(N)}
lsdc = log10(sdc[-1])
logN = log10(2:100)
plot(logN,lsdc,type="l",
+ col="red")
```



1. The benefit from averaging
   - The Law of Large Numbers
   - Estimate distribution parameters
   - Howto reduce noise – Graphical illustration
2. Convergence of the averaging
   - Convergence of the mean for a standard Gaussian
   - And if there are potential outliers?
   - Non convergence of the average for a Cauchy
3. Comparing two empiric means
   - Robustness of the *t* statistic
   - Estimate the t Statistics
   - Monte Carlo simulation of the H_0 rejection

The *t* statistic

Suppose we have two independent samples $x_1, x_2, ..., x_m$ and $y_1, y_2, ..., y_n$ from statistical variables X and Y, and we wish to test the null hypothesis $H_0 : \mu_X == \mu_Y$ that the actual means of both variables X and Y are in fact the same.
The standard test for this H_0 is based on the t-statistic:

$$T = \frac{\overline{x} - \overline{y}}{\sigma_p \sqrt{(1/m + 1/n)}}$$

where $\overline{x}$ and $\overline{y}$ are the resepective observed sample means, and σ_p is the *pooled* standard deviation:

$$\sigma_p = \sqrt{\frac{(m-1)\sigma_x^2 + (n-1)\sigma_y^2}{m+n-2}}$$

Robustness of the t Statistics

Under H_0, $T \sim \mathcal{T}(df = m + n - 2)$. Suppose the level of signficance of the test is set at α, then one rejects H_0 when $|T| \geqslant t_{n+m-2,\alpha/2}$ where $t_{df,p}$ is the $1 - p$ quantile of a t random variable with df degrees of freedom.

The underlying assumptions of the test are:

1. X and Y are independent normal distributed variables,
2. X and Y admit the same variance.

An interesting problem is to investigate the robustness of this popular test with respect to changes in the assumptions.

Writing a function to estimate the t Statistic

Here some R commands for computing a t statistic:

```
> X = rnorm(10,mean=50,sd=10)
> Y = rnorm(10,mean=50,sd=10)
> m = length(X)
> n = length(Y)
> sp = sqrt(( (m-1)*sd(X)^2 +
+   (n-1)*sd(Y)^2 ) / ( m+n -2) )
> t = ( mean(X) - mean(Y) ) /
+   ( sp * sqrt(1/m + 1/n) )
```

We may write a R function tstatistic to compute these results in the future.

The following text is saved in file “tstatistics.R”:

```
tstatistic = function(X,Y)
{
m = length(X)
n = length(Y)
sp = sqrt(( (m-1)*sd(X)^2 +
  (n-1)*sd(Y)^2 ) / ( m+n -2) )
t = ( mean(X) - mean(Y) ) /
  ( sp * sqrt(1/m + 1/n) )
return(t)
}
```

We may load this function in R with the command > source("tstatistic.R").

true significance of H_0 rejection test

True significance of the t statistic will depend on:

- the required α level of significance of the test,
- the shape of the distributions X and Y,
- the spreads of the distributions X and Y, and
- the sample sizes m and n.

Monte Carlo simluation of the H_0 rejection

Given a particular choice of α, shape, spreads, and sample sizes, we wish to estimate the true signficance level of the H_0 rejection test given by:

$$\alpha^T = P(|T| \geqslant t_{n+m-2,\alpha/2})$$

Here an outline of a simulation algorithm: Repeat $nSim$ times:

1. generate independent sequences of the X and Y random variables,
2. compute the empirical T statistic from the two samples,
3. if $|T|$ exceeds the theoretical t value, reject H_0

The estimate $\hat{\alpha}^T$ of the true significance is given as the ratio of the number of rejections of H_0 over $nSim$.

Robustness of the true significance level

Exercise

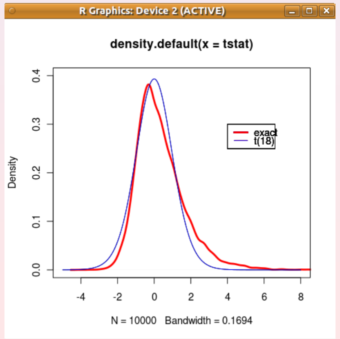
Suppose we fix the required significance level at $\alpha = 0.1$ and keep the sample sizes at $m = 10$ and $n = 10$. One may simulate $nSim = 10^4$ t-statistics with the following assumptions:

1. *normal Z variables (zero means and spreads of one)*
2. *normal variables with zero means and spreads of one, respectively 10,*
3. *T variables with 4 dfs and equal spreads,*
4. *exponential variables with equal mean of one,*
5. *one normal variable (mean=10, sd=2) and one exponential variable with mean = 10.*

Graphical illustration of the T distributions

We may illustrate the empirical T distribution for instance in the fourth case where we suppose a normal and an exponential variable.
We suppose that the `nSim` simulated values of the `t` statistic are gathered in a `tstat` vector:

```
> tstat = rep(0,nSim)
> for (i in 1:nSim){
+       X=rnorm(10,mean=10,sd=2)
+       Y=rexp(10,rate=1/10)
+       tstat[i] = tstatistic(X,Y) }
> plot(density(tstat),xlim=c(-5,8),
+   ylim=c(0,.4), lwd=3, col="red")
> x = seq(-5,8,length=200)
> lines(x,dt(x,df=18),col="blue")
> legend(4,.3,c("exact","t(18)"),
+   lwd=c(3,1), col=c("red","blue"))
```



Exercise

Exercise (Robustness of the confidence interval of proportions)

Suppose one observes a random variable X that is supposed to be binomially distributed with a sample size n and a success probability of p. The standard 90% confidence interval of p is given by

$$C(X) \;=\; \left[\hat{p} - z_{0.9}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{0.9}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

where $\hat{p} = \sum X/n$.
We rely in this approach on the assumption that $P(p \in C(X)) = 0.90$ for all $0 < p < 1$.

Exercise – continue

Exercise (Robustness of the confidence interval of proportions)

Questions:

1. *Write a R-function called `binomialConfInterval` that returns the limits of a 90% confidence interval for a simulation of a binomial random variable X with sample size n.*
2. *Simulate $nSim = 1000$ times the computation of the confidence interval when $n = 20$ and the true value of p is 0.5 and estimate the true probability of coverage.*
3. *Construct a Monte Carlo study that investigates how the probability of coverage depends on the sample size n and true proportion value. Let n take the values 10,25, and 50 and let p be 5%, 25%, and 50%. The number of simulations nSim be 1000 in each case.*
4. *Write a function that takes three arguments: n, p and nSim, and returns the estimate of the true coverage probability.*
5. *Describe how the actual coverage probability of the confidence interval estimate depends in fact on the sample size and true success proportion of the underlying binomial process.*