Theory supporting the net benefit and Peron's scoring rules

Brice Ozenne

December 12, 2018

This document describe the relationship between the net benefit and traditional parameter of interest (e.g. hazard ratio). It also present how Peron's scoring rules for the survival and competing setting were derived.

In the examples we will use a sample size of:

```
n <- 1e4
```

and use the following R packages

```
library(BuyseTest)
library(riskRegression)
library(survival)
```

Contents

1 References

Lee, A. J. (1990). U-statistics: Theory and practice. statistics: Textbooks and monographs 110. *Dekker, New York. MR*, 10754:17.

A Recall on the U-statistic theory

This recall is based on chapter 1 of Lee (1990).

A.1 Motivating example

We will illustrate basic results on U-statistics with the following motivating question: "what is the asymptotic distribution of the empirical variance estimator?". For a more concrete example, imagine that we want to provide an estimate with its 95% confidence interval for the variability in cholesterol measurements. We assume that we are able to collect a sample of n independent and identically distributed (iid) realisations $(x_1, \ldots, x_n)$ of the random variable cholesterol, denoted X. We ignore any measurement error.

A.2 Estimate, estimator, and functionnal

We can compute an **estimate** s of the variance using the following **estimators** f_m and f_s:

$$m = f_m(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$s = f_s(x_1, \ldots, x_n) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 \tag{2}$$

Given a dataset s is a deterministic (not random) quantity, e.g. if we observe:

```
x <- c(1,3,5,2,1,3)
```

then s equals:

```
m <- mean(x)
s <- var(x)
s
```

```
[1] 2.3
```

But in general the value of s depends on the dataset. s is like a function f_n that takes as argument some data and output a quantity of interest. This is often refer to as a **functionnal**, e.g. $S = f_n(x_1, \ldots, x_n)$. Here we use the upper case S to emphasise that it is a random quantity: for each new realisation $(x_1, \ldots, x_n)$ of X corresponds a realisation for S i.e. a possibly different value for the variance. If mechanism generating the data has cumulative distribution function F then we can also define the true value as $\mu = f(F) = lim$. Because S and f are very close quantities we will not distinguish them in the notation, i.e. write $S = S(F)$. This corresponds to formula (1) in Lee (1990). To be more precise we can explicit what is $S(F)$:

$$M(F) = \int_{-\infty}^{+\infty} x dF(x) \tag{3}$$

$$S(F) = \int_{-\infty}^{+\infty} (x - M(F))^2 dF(x) \tag{4}$$

When we observe a sample, we use it to plug-in formula (3) and (4) an approximation $\hat{F}$ of F. Usually our best guess for F is $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x \leq x_i}$ where $\mathbb{1}_{.}$ is the indicator function taking value 1 if . is true and 0 otherwise. One can check that when plug-in $\hat{F}$ formula (3) and (4) becomes formula (1) and (2).

To summarize:

- an estimator is a random variable whose realisation depends on the data. It realization is called estimate.
- an estimate is a deterministic value that we obtain using the observed data (e.g. variability is 2.3)
- a functionnal (of an estimator) is the rule by which an estimator transform the data into an estimate.

A.3 Aim

Using formula (1) and (2) we can easily estimate the variance based on the observed realisations of X (i.e. the data). However how can we get an confidence interval? What we want is to quantify the incertainty associated with the estimator, i.e. how the value output by the functionnal is sensitive to a change in the dataset. To do so, since the estimator S is a random variable, we can try to characterize its distribution. This is in general difficult. It is much easier to look at the distribution of the estimator S if we would have an infinite sample size. This is what we will do, and rely on similations to see how things go in finite sample size. As we will see, the asymptotic distribution of the variance is a Gaussian distribution with a variance that we can estimate:

```
n <- length(x)
k <- mean((x-m)^4)
var_s <- k/n - ((n-3)*s^2)/(n*(n-1))
var_s
```

```
[1] 0.8425278
```

So we obtain a 95% confidence intervals for the variance doing:

```
c(estimate = s,
  lower = s + qnorm(0.025) * sqrt(var_s),
  upper = s + qnorm(0.975) * sqrt(var_s))
```

```
 estimate     lower     upper
2.3000000 0.5009625 4.0990375
```

We can see that it is not a very good confidence interval since it symmetric - we know that the variance is positive so it should extend more on the right side. But this only problematic in small sample sizes. In large enough sample sizes the confidence interval will be correct and we focus on this case.

In summary, we would like:

- to show that our estimator S is asymptotically normally distributed.
- to have a formula for computing the asymptotic variance

Note: we can already guess that the estimator S (as most estimators) will be asymptotically distributed because it can be expressed as a average (see formula (2)). If we would know the mean of X, then the terms $x_i - m$ are iid so the asymptotically normality of S follows from the central limit theorem. It does not give us a formula for the asymptotic variance though. To get that we will use results from the theory of the U-statistics.

A.4 Definition of a U-statistic and examples

A U-statistic with kernel h of order k is an estimator of the form:

$$U_n = \frac{1}{\binom{n}{k}} \sum_{(\beta_1,\ldots,\beta_r)\in\beta} h\left(X_{\beta_1},\ldots,X_{\beta_k}\right)$$

where *beta* is the set of all possible permutations between r integers choosen from $\{1,\ldots,n\}$. The simplest example of a U-statistic is the estimator of mean for which $k = 1$ and h is the identity function:

$$U_n = \frac{1}{\binom{n}{1}} \sum_{(\beta_1)\in\{1,\ldots,n\}} X_{\beta_1} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Our estimator of the variance is also a U-statistic, but this requires a little bit more work to see that:

$$\begin{aligned}
s &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i m + m^2\right) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\frac{1}{n}\sum_{j=1}^{n}x_j + m^2\right) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i^2 - 2x_i x_j + m^2\right) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((x_i - x_j)^2 - x_j^2 + m^2\right) \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 - \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j^2 - m^2\right) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2 - s \\
s &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{(x_i - x_j)^2}{2} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{i<j}^{n}\frac{(x_i - x_j)^2}{2} = \frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{i<j}^{n}\frac{(x_i - x_j)^2}{2}
\end{aligned}$$

So the variance estimator is a U-statistic of order 2 with kernel $k(x) = (x - m)^2$.

A.5 Why using the U-statistic theory?

Because it provides very useful results to characterize the asymptotic distribution of an estimator. For instance:

Theorem 3 (Lee (1990)) Let U_n be a U-statistic with a kernel h of degre r. Then:

$$\mathbb{V}ar\left[U_n\right] = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\sigma_c^2$$
$$\text{where } \sigma_c^2 = \mathbb{C}ov\left[h(x_1,\ldots,x_r), h(x'_1,\ldots,x'_r)\right]$$