Theory supporting the net benefit and Peron's scoring rules

Brice Ozenne

December 17, 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 $(x_1, \dots, x_n)$ of the random variable cholesterol, denoted X. We ignore any measurement error.

A.2 Estimate, estimator, and functionnal

We can compute an **estimate** of the variance using the following **estimators** $\hat{\mu}$ and $\hat{\sigma}^2$:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \tag{2}$$

Given a dataset the estimator $\hat{\sigma}^2$ outputs a deterministic (i.e. not random) quantity, called the estimate of the variance. For instance if we observe:

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

then s equals:

```
mu <- mean(x)
sigma2 <- var(x)
sigma2
```

```
[1] 2.3
```

In general the value of the estimate depends on the dataset. The estimator acts 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. $\hat{\sigma}^2 = f_n(x_1, \dots, x_n)$. Here we use the hat notation to emphasise that $\hat{\sigma}^2$ is a random quantity: for each new realisation $(x_1, \dots, x_n)$ of X corresponds a realisation for $\hat{\sigma}^2$ 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 $\sigma^2 = f_{\sigma^2}(F)$ (which is a deterministic value) where:

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

$$\sigma^2(F) = f_{\sigma^2}(F) = \int_{-\infty}^{+\infty} (x - f_\mu(F))^2 dF(x) \tag{4}$$

This can be understood as the limit $f(F) = \lim_{n\to\infty} f_n(x_1, \ldots, x_n)$. Because σ^2 and f_{σ^2} are very close quantities we will not distinguish them in the notation, i.e. write $\sigma^2 = \sigma^2(F)$. This corresponds to formula (1) in Lee (1990).

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. Its realization is called estimate.
- an estimate is a deterministic value that we obtain using the observed data (e.g. observed variability is 2.3)
- a functionnal (of an estimator) is the rule by which an estimator transforms 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 $\hat{\sigma}^2$ 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 $\hat{\sigma}^2$ 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-mu)^4)
var_sigma2 <- (k-sigma2^2)/n
var_sigma2
```

```
[1] 0.4898611
```

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

```
c(estimate = sigma2,
  lower = sigma2 + qnorm(0.025) * sqrt(var_sigma2),
  upper = sigma2 + qnorm(0.975) * sqrt(var_sigma2))
```

```
 estimate     lower     upper
2.3000000 0.9282197 3.6717803
```

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 $\hat{\sigma}^2$ is asymptotically normally distributed.
- to have a formula for computing the asymptotic variance.

To do so we will use results from the theory on U-statistics.

NOTE: we can already guess that the estimator $\hat{\sigma}^2$ (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 $x_i - \mu$ are iid so the asymptotically normality of $\hat{\sigma}^2$ follows from the central limit theorem. It does not give us a formula for the asymptotic variance though.

A.4 Definition of a U-statistic and examples

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

$$\hat{U} = \frac{1}{\binom{n}{k}} \sum_{(\beta_1,\ldots,\beta_k)\in\beta} h\left(x_{\beta_1},\ldots,x_{\beta_k}\right)$$

where β is the set of all possible permutations between k integers choosen from $\{1,\ldots,n\}$. We will also assume that the kernel is symmetric, i.e. the order of the arguments in h has no importance. Note that because the observations are iid, $\hat{U}$ is an unbiased estimator of U.

EXAMPLE 1: the simplest example of a U-statistic is the estimator of mean for which $k = 1$ and h is the identity function:

$$\hat{\mu} = \frac{1}{\binom{n}{1}} \sum_{(\beta_1)\in\{1,\ldots,n\}} x_{\beta_1} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

EXAMPLE 2: our estimator of the variance is also a U-statistic, but this requires a little bit more work to see that:

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

So the variance estimator is a U-statistic of order 2 with kernel $h(x_1, x_2) = \frac{(x_1-x_2)^2}{2}$.

EXAMPLE 3: another classical example of U-statistic is the signed rank statistic which enable to test non-parametrically whether the center of a distribution is 0. This corresponds to:

```
wilcox.test(x)
```

```
        Wilcoxon signed rank test with continuity correction

data:  x
V = 21, p-value = 0.03501
alternative hypothesis: true location is not equal to 0

Warning message:
In wilcox.test.default(x) : cannot compute exact p-value with ties
```

Let's take to random realisation of X and denote thoses X_1 and X_2 (they are random variables). The parameter of interest (or true value) is $U = \mathbb{P}[X_1 + X_2 > 0]$ and the corresponding estimator is:

$$\hat{U} = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n} \sum_{i<j} \mathbb{1}_{x_i+x_j>0}$$

A.5 A major result from the U-statistic theory

So far we have seen that our estimator for the variance was a U-statistic. We will now use the U-statistic theory to obtain its asymptotic distribution.

Theorem (adapted from Lee (1990), theorem 1 page 76)
Let $\hat{U}$ be a U-statistic of order k with non-zero first component in its H-decomposition. Then $n^{\frac{1}{2}}(\hat{U} - U)$ is asymptotically normal with mean zero and asymptotic variance σ_1^2 where σ_1^2 is the variance of the first component in the H-decomposition of $\hat{U}$.

So under the asymption that the first term of the H-decomposition of the variance is non 0 then we know that the asymptotic distribution of our variance estimator is normal and if we are able to compute the variance of the first term of the H-decomposition then we would also know the variance parameter of the asymptotic distribution. So it remains to see what is this H-decomposition and how can we characterize it.

A.6 The first term of the H-decomposition

The H-decomposition (short for Hoeffling decomposition) enables us to decompose the estimator of a U-statistic of rank k into a sum of k uncorrelated U-statistics of increasing order (from 1 to k) with variances of decreasing order in n. As a consequence the variance of the U-statistic will be asymptotically equal to the variance of the first non-0 term in the decomposition.

Before going further we introduce:

- $X_1, \ldots, X_n$ the random variables associated with each sample.
- $\mathcal{L}_2$ the space of all random variables with zero mean and finite variance.
  It is equiped with the inner product $\mathbb{C}ov\left[X, Y\right]$.
- the subspaces $\left(\mathcal{L}_2^{(j)}\right)_{j\in\{1,\ldots,k\}}$ where for a given $j \in \{1, \ldots, k\}$, $\mathcal{L}_2^{(j)}$ is the subspace of $\mathcal{L}_2$ containing all random variables of the form $\sum_{(\beta_1,\ldots,\beta_j)\in\beta} \psi(X_{\beta_1}, \ldots, X_{\beta_j})$ where β is the set of all possible permutations between j integers choosen from $\{1, \ldots, n\}$. For instance $\mathcal{L}_2^{(1)}$ contains the mean, $\mathcal{L}_2^{(2)}$ contains the variance, and $\mathcal{L}_2^{(j)}$ contains all U-statistics of order j with square integrable kernels.

We can now define the H-decomposition as the projection of $\hat{U} - U$ on the subspaces $\mathcal{L}_2^{(1)}$, $\mathcal{L}_2^{(2)} \cap \left(\mathcal{L}_2^{(1)}\right)^{\perp}$, $\ldots$, $\mathcal{L}_2^{(k)} \cap \left(\mathcal{L}_2^{(k-1)}\right)^{\perp}$. Here $A^{\perp}$ indicates the space orthogonal to A. So the first term of the H-decomposition, denoted $H^{(1)}$, is the projection of $\hat{U} - U$ on $\mathcal{L}_2^{(1)}$; this is also called the Hájek projection. Clearly all terms of the projection are mutually orthogonal (or uncorrelated), they are unique (it is a projection) and they correspond to U-statistics of increasing degree (from 1 to k). It remains to get a more explicit expression for these term and show that their variance are of decreasing order in n.

We now focus on the first term and show that $H^{(1)} = \sum_{i=1}^n \mathbb{E}\left[\hat{U} - U|X_i\right]$. Clearly this term belongs to $\mathcal{L}_2^{(1)}$. It remains to show that $\hat{U} - U - H^{(1)}$ is orthogonal to $\mathcal{L}_2^{(1)}$. Let consider an element $V \in \mathcal{L}_2^{(1)}$:

$$\begin{aligned}
\mathbb{C}ov\left[\hat{U} - U - H^{(1)}, V\right] &= \mathbb{E}\left[(\hat{U} - U - H^{(1)})V\right] \\
&= \sum_{i'=1}^n \mathbb{E}\left[(\hat{U} - U - H^{(1)})\psi(X_{i'})\right] \\
&= \sum_{i'=1}^n \mathbb{E}\left[\mathbb{E}\left[\hat{U} - U - H^{(1)}\middle|X_{i'}\right]\psi(X_{i'})\right]
\end{aligned}$$

So it remains to show that $\mathbb{E}\left[\hat{U} - U\middle|X_{i'}\right] = \mathbb{E}\left[H^{(1)}\middle|X_{i'}\right]$. This follows from:

$$\begin{aligned}
\mathbb{E}\left[H^{(1)}\middle|X_{i'}\right] &= \mathbb{E}\left[\sum_{i=1}^n \mathbb{E}\left[\hat{U} - U|X_i\right]\middle|X_{i'}\right] = \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left[\hat{U} - U|X_i\right]\middle|X_{i'}\right] \\
&= \mathbb{E}\left[\hat{U} - U|X_i\right] + \sum_{i\neq i'}^n \mathbb{E}\left[\mathbb{E}\left[\hat{U} - U|X_i\right]\middle|X_{i'}\right] \\
&= \mathbb{E}\left[\hat{U} - U|X_i\right] + \sum_{i\neq i'}^n \mathbb{E}\left[\mathbb{E}\left[\hat{U} - U|X_i\right]\right]
\end{aligned}$$

where we have used that X_i and $X_{i'}$ are independent and $\mathbb{E}\left[\mathbb{E}\left[\hat{U} - U|X_i\right]\right] = \mathbb{E}\left[\hat{U} - U\right] = 0$.

We can now re-express the first term of the H-decomposition more explicitely:

$$
\begin{aligned}
H^{(1)} &= \sum_{i=1}^{n} \mathbb{E}\left[\hat{U} - U \middle| X_i\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{\binom{n}{k}} \sum_{(\beta_1,\ldots,\beta_k)\in\beta} h\left(x_{\beta_1},\ldots,x_{\beta_k}\right) - U \middle| X_i\right] \\
&= \frac{1}{\binom{n}{k}} \sum_{(\beta_1,\ldots,\beta_k)\in\beta} \sum_{i=1}^{n} \mathbb{E}\left[h\left(x_{\beta_1},\ldots,x_{\beta_k}\right) \middle| X_i\right] - U \\
&= \frac{1}{\binom{n}{k}} \sum_{(\beta_1,\ldots,\beta_k)\in\beta} \sum_{i=1}^{n} \mathbb{1}_{i\in\beta}\mathbb{E}\left[h\left(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_k\right) \middle| x_i\right] + \mathbb{1}_{i\notin\beta} * 0 - U \\
&= \frac{1}{\binom{n}{k}} \sum_{i=1}^{n} \mathbb{P}\left[i\in\beta\right] \mathbb{E}\left[h\left(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_k\right) \middle| x_i\right] - U \\
&= \frac{\binom{n-1}{k-1}}{\binom{n}{k}} \sum_{i=1}^{n} \mathbb{E}\left[h\left(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_k\right) \middle| x_i\right] - U \\
H^{(1)} &= \frac{k}{n} \sum_{i=1}^{n} \mathbb{E}\left[h\left(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_k\right) \middle| x_i\right] - U
\end{aligned}
$$

Let's now compute the variance of $\hat{U}$:

$$
\begin{aligned}
\mathbb{V}ar\left[\hat{U}\right] &= \binom{n}{k}^{-2} \mathbb{V}ar\left[\sum_{(\beta_1,\ldots,\beta_k)\in\beta} h\left(x_{\beta_1},\ldots,x_{\beta_k}\right)\right] \\
&= \binom{n}{k}^{-2} \mathbb{C}ov\left[\sum_{(\beta_1,\ldots,\beta_k)\in\beta} h\left(x_{\beta_1},\ldots,x_{\beta_k}\right), \sum_{(\beta'_1,\ldots,\beta'_k)\in\beta'} h\left(x_{\beta'_1},\ldots,x_{\beta'_k}\right)\right] \\
&= \binom{n}{k}^{-2} \sum_{(\beta_1,\ldots,\beta_k)\in\beta} \sum_{(\beta'_1,\ldots,\beta'_k)\in\beta'} \mathbb{C}ov\left[h\left(x_{\beta_1},\ldots,x_{\beta_k}\right), h\left(x_{\beta'_1},\ldots,x_{\beta'_k}\right)\right]
\end{aligned}
$$

Using the symmetry of the kernel we see that the terms in the double sum only depends on the number of common observations. To determine a term with j common observations, a choose:

- k observations among the n for the first kernel: $\binom{n}{k}$ possibilities
- c common index for the two kernels among the k: $\binom{k}{c}$ possibilities
- $k-c$ observations among the remaining $n-k$ observations for the second kernel: $\binom{n-k}{k-c}$ possibilities

So denoting $\sigma_c^2 = \mathbb{C}ov\left[h\left(x_1, \ldots, x_k\right), h\left(x_1, \ldots, x_c, x'_{c+1}, \ldots, x'_k\right)\right]$ this gives:

$$\begin{aligned}
\mathbb{V}ar\left[\hat{U}\right] &= \binom{n}{k}^{-2} \sum_{c=0}^{n} \binom{n}{k}\binom{k}{c}\binom{n-k}{k-c}\sigma_c^2 \\
&= \sum_{c=0}^{k} \frac{k!(n-k)!}{n!} \frac{k!}{c!(k-c)!} \frac{(n-k)!}{(k-2k+c)!(n-c)!}\sigma_c^2 \\
&= \sum_{c=0}^{k} \frac{k!^2}{c!(k-c)!^2} \frac{(n-k)!^2}{(n-2k+c)!n!}\sigma_c^2 \\
&= \sum_{c=0}^{k} \mathcal{O}\left(\frac{(n-k)!^2}{(n-2k+c)!n!}\right)\sigma_c^2 \\
&= \sum_{c=0}^{k} \mathcal{O}\left(\frac{(n-k)\ldots(n-2k+c+1)}{n\ldots(n-k+1)}\right)\sigma_c^2 \\
&= \sum_{c=0}^{k} \mathcal{O}\left(\frac{n^{-k+2k-c}}{n^k}\right) = \sum_{c=0}^{k} \mathcal{O}\left(n^{-c}\right)\sigma_c^2
\end{aligned}$$

So if $\sigma_1^2 \neq 0$ then the asymptotic variance only depends on the variance of the first term, i.e.:

$$\begin{aligned}
\mathbb{V}ar\left[\hat{U}\right] = \mathbb{V}ar\left[H^{(1)}\right] &= \frac{k^2}{n^2}\mathbb{V}ar\left[\sum_{i=1}^{n}\mathbb{E}\left[h\left(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k\right)\middle|x_i\right]\right] \\
&= \frac{k^2}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left[\mathbb{E}\left[h\left(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_k\right)\middle|x_i\right]\right] \\
&= \frac{k^2}{n^2}n\mathbb{V}ar\left[\mathbb{E}\left[h\left(x, x_2, \ldots, x_k\right)\middle|x\right]\right] \\
\mathbb{V}ar\left[\hat{U}\right] &= \frac{k^2}{n}\mathbb{V}ar\left[\mathbb{E}\left[h\left(x, x_2, \ldots, x_k\right)\middle|x\right]\right]
\end{aligned}$$

In summary we have obtained a formula for the asymptotic variance of the U-statistic.

Example 1: Sample mean

We first compute the Hájek projection of the mean:

$$H_{\hat{\mu}}^{(1)} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i|x_i\right] - \mu = \frac{1}{n}\sum_{i=1}^{n}x_i - \mu$$

And then compute the asymptotic variance as:

$$\mathbb{V}ar\left[\hat{\mu}\right] = \mathbb{V}ar\left[H_{\hat{\mu}}^{(1)}\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left[x_i - \mu\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}$$

Example 2: Sample variance

We first compute the Hájek projection of the variance:

$$\begin{aligned} H_{\hat{\sigma}^2}^{(1)} &= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{(x_i - X_2)^2}{2}\Big|x\right] - \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i^2 - 2x_iX_2 + X_2^2\Big|x_i\right] - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2x_i\mu + \sigma^2 + \mu^2\right) - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left((x_i - \mu)^2 - \sigma^2\right) \end{aligned}$$

And then compute the asymptotic variance as:

$$\begin{aligned} \mathbb{V}ar\left[\hat{\sigma}^2\right] = \mathbb{V}ar\left[H_{\hat{\sigma}^2}^{(1)}\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left[(x_i - \mu)^2 - \sigma^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[(x-\mu)^4\right] - \mathbb{E}\left[(x-\mu)^2\right]^2 \\ &= \frac{\mu_4 - (\sigma^2)^2}{n} \end{aligned}$$

where $\mu_4 = \mathbb{E}\left[(x-\mu)^4\right]$ is the fourth moment of the distribution. For a better approximation in small sample size we could account for the variance of the second term of the H-decomposition. We would obtain (Lee (1990), page 13):

$$\mathbb{V}ar\left[\hat{\sigma}^2\right] = \frac{\mu_4}{n} - \frac{(n-3)(\sigma^2)^2}{n(n-1)}$$

When $\frac{n-3}{n-1}$ is close to 1 then the first order approximation is sufficient.

Example 3: Signed rank statistic
We first compute the Hájek projection of the signed rank statistic:

$$\begin{aligned} H_{\hat{U}}^{(1)} &= \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbb{1}_{x_i+X_2>0}\Big|x_i\right] - U = \frac{2}{n}\sum_{i=1}^{n}\mathbb{P}\left[X_2 > -x_i\Big|x_i\right] - \mathbb{P}\left[X_2 > -X_1\right] \\ &= \frac{2}{n}\sum_{i=1}^{n}(1 - F(-x_i)) - \mathbb{E}_x\left[(1 - F(-x))\right] \end{aligned}$$

Since under the null, the distribution is symmetric $F(-x) = 1 - F(x)$:

$$H_{\hat{U}}^{(1)} = \frac{2}{n}\sum_{i=1}^{n}F(x_i) - \mathbb{E}_x\left[F(x)\right]$$

We will use that for continuous distribution $F(x)$ is uniformly distribution and therefore has variance $\frac{1}{12}$. So we can compute the asymptotic variance as:

$$\mathbb{V}ar\left[\hat{U}\right] = \mathbb{V}ar\left[H_U^{(1)}\right] = \frac{4}{n^2}\sum_{i=1}^{n}\mathbb{V}ar\left[F(x_i) - \mathbb{E}_x\left[F(x)\right]\right] = \frac{4}{n^2}n\frac{1}{12} = \frac{1}{3}$$

A.7 Two sample U-statistics

So far we have assumed that all our observations were iid. But in the case of GPC, we study two populations (experimental arm and control arm) so we can only assume to have two independent samples $x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$ where the first one contains iid realisations of a random variable X and the second one contains iid realisations of a second variable Y. We can now define a two-sample U-statistic as of order k_x and k_y as:

$$\hat{U} = \frac{1}{\binom{m}{k_x}\binom{n}{k_y}} \sum_{(\alpha_1,\ldots,\alpha_{k_x})\in\alpha} \sum_{(\beta_1,\ldots,\beta_{k_y})\in\alpha} h(x_{\alpha_{k_x}}, \ldots, x_{\alpha_j}, y_{\beta_1}, \ldots, y_{\beta_{k_y}})$$

where α (resp. β) is the set of all possible permutations between k_x (resp. k_y) intergers chosen from $\{1, \ldots, m\}$ (resp. $\{1, \ldots, n\}$) and the kernel $h = h(x_1, \ldots, x_{k_x}, y_1, \ldots, y_{k_y})$ is permutation symmetric in its first k_x arguments and its last k_y arguments separately. Once more it follows from the independence and iid assumptions that $\hat{U}$ is an unbiased estimator of $U = \mathbb{E}\left[h(X_1, \ldots, X_{k_x}, Y_1, \ldots, Y_{k_y})\right]$ where $X_1, \ldots, X_{k_x}$ (resp. $Y_1, \ldots, Y_{k_y}$) are the random variables associated to distinct random samples from X (resp. Y). The two-sample case is a specific case of the Generalized U-statistics introduced in section 2.2 in Lee (1990).

Many results for U-statistics extends to two sample U-statistics. For instance the Hájek projection of $\hat{U} - U$ becomes:

$$H^{(1)} = \frac{k_x}{m}\sum_{i=1}^{m}\left(\mathbb{E}\left[h(x, x_2, \ldots, x_{k_x}, y_1, \ldots, y_{k_y})\middle|x\right] - U\right) + \frac{k_y}{n}\sum_{j=1}^{n}\left(\mathbb{E}\left[h(x_1, \ldots, x_{k_x}, y, y_2, \ldots, y_{k_y})\middle|y\right] - U\right)$$

Before stating any asymptotic results, we need to define what we now mean by asymptotic (since we have two sample sizes m and n). We now mean by asymptotic that we create an increasing sequence of m and n indexed by v such that:

- $m_v \xrightarrow[v\to\infty]{} \infty$
- $n_v \xrightarrow[v\to\infty]{} \infty$
- there exist a $p \in]0;1[$ satisfying $\frac{m}{n+m} \xrightarrow[v\to\infty]{} p$ and $\frac{n}{n+m} \xrightarrow[v\to\infty]{} 1-p$.

Informally speaking, this means that m and n goes to infinity at the same speed. Let's denotes:

$$\mathbb{V}ar\left[\mathbb{E}\left[h(x, x_2, \ldots, x_{k_x}, y_1, \ldots, y_{k_y})\middle|x\right]\right] = \sigma^2_{1,0}$$
$$\mathbb{V}ar\left[\mathbb{E}\left[h(x_1, \ldots, x_{k_x}, y, y_2, \ldots, y_{k_y})\middle|y\right]\right] = \sigma^2_{0,1}$$

We then have the following result:

Theorem (adapted from Lee (1990), theorem 1 page 141)
Let $\hat{U}$ be a U-statistic of order k_x and k_y with non-zero first component (i.e. $\sigma^2_{1,0} > 0$ and $\sigma^2_{0,1} > 0$) in its H-decomposition. Then $(m+n)^{\frac{1}{2}}(\hat{U}-U)$ is asymptotically normal with mean zero and asymptotic variance $p^{-1}k_x^2\sigma^2_{1,0} + (1-p)^{-1}k_y^2\sigma^2_{0,1}$ which is the variance of the first component in the H-decomposition of $\hat{U}$.

Example 4: Mann-Whitney statistic

If our parameter of interest is $\mathbb{P}[X \leq Y]$ then the estimator:

$$\hat{U} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{1}_{x_i \leq y_j}$$

is a U-statistic of order $k_x = 1$ and $k_y = 1$ with kernel $h(x, y) = \mathbb{1}_{x \leq y}$ We first compute the Hájek projection of the signed rank statistic:

$$\begin{aligned}
H_{\hat{U}}^{(1)} &= \frac{1}{m} \sum_{i=1}^{m} \left(\mathbb{E}\left[\mathbb{1}_{x_i \leq y} \middle| x_i \right] - U \right) + \frac{1}{n} \sum_{j=1}^{n} \left(\mathbb{E}\left[\mathbb{1}_{x \leq y_j} \middle| y_j \right] - U \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(\mathbb{P}[Y \geq x_i] - U \right) + \frac{1}{n} \sum_{j=1}^{n} \left(\mathbb{P}[X \leq y_j] - U \right) \\
&= \frac{1}{m} \sum_{i=1}^{m} \left(1 - F_{-,y}(x_i) - U \right) + \frac{1}{n} \sum_{j=1}^{n} \left(F_x(y_j) - U \right) \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left(F_{-,y}(x_i) - \mathbb{E}_x[F_{-,x}(x)] \right) + \frac{1}{n} \sum_{j=1}^{n} \left(F_x(y_j) - \mathbb{E}_y[F_y(y)] \right)
\end{aligned}$$

where F_- is the left limit of F, F_x(resp. F_y) denoting the cumulative distribution function of X (resp. Y). For continuous distributions $F_- = F$ and under the null hypothesis that $F_x = F_y$, we get that:

$$\mathbb{V}ar\left[\hat{U}\right] = \mathbb{V}ar\left[H_{\hat{U}}^{(1)}\right] = \frac{1}{m}\frac{1}{12} + \frac{1}{n}\frac{1}{12} = \frac{nm}{12(m+n)}$$

If we are not under the null we end up with the formula:

$$\mathbb{V}ar\left[\hat{U}\right] = \frac{1}{m^2} \sum_{i=1}^{m} \mathbb{V}ar\left[\mathbb{E}\left[\mathbb{1}_{x_i \leq y} \middle| x_i \right] - U \right] + \frac{1}{n^2} \sum_{j=1}^{n} \mathbb{V}ar\left[\mathbb{E}\left[\mathbb{1}_{x \leq y_j} \middle| y_j \right] - U \right]$$

Noticing that:

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{x_i \leq y} \middle| x_i \right] - U \right] = \mathbb{E}\left[\mathbb{1}_{x_i \leq y} \right] - U = 0$$

We can compute the variance as:

$$\begin{aligned}
\mathbb{V}ar\left[\mathbb{E}\left[\mathbb{1}_{x_i \leq y} \middle| x_i \right] - U \right] &= \mathbb{E}\left[\left(\mathbb{E}\left[\mathbb{1}_{x_i \leq y} \middle| x_i \right] - U \right)^2 \right] \\
&= \int_x \left(\int_y \left(\mathbb{1}_{x \leq y} - U \right) dF_Y(y) \right) \left(\int_y \left(\mathbb{1}_{x \leq y} - U \right) dF_Y(y) \right) dF_X(x) \\
&= \int_x \left(\int_{y_1} \left(\mathbb{1}_{x \leq y_1} - U \right) dF_Y(y_1) \right) \left(\int_{y_2} \left(\mathbb{1}_{x \leq y_2} - U \right) dF_Y(y_2) \right) dF_X(x) \\
&= \mathbb{E}\left[\left(\mathbb{1}_{x \leq y_1} - U \right) \left(\mathbb{1}_{x \leq y_2} - U \right) \right] \\
&= \mathbb{E}\left[\mathbb{1}_{x \leq x_1} \mathbb{1}_{x \leq y_2} \right] - \mathbb{E}\left[\mathbb{1}_{x \leq y_1} \right] U - \mathbb{E}\left[\mathbb{1}_{x \leq y_2} \right] U + U^2 \\
&= \mathbb{P}[x \leq y_1, x \leq y_2] - \mathbb{P}[x \leq y]^2
\end{aligned}$$

So the variance is:

$$\begin{aligned}\mathbb{V}ar\left[\hat{U}\right] &= \frac{1}{m}\left(\mathbb{P}\left[x \leq y_1, x \leq y_2\right] - \mathbb{P}\left[x \leq y\right]^2\right) + \frac{1}{n}\left(\mathbb{P}\left[x_1 \leq y, x_2 \leq y\right] - \mathbb{P}\left[x \leq y\right]^2\right) \\ &= \frac{\sigma_{1,0}^2}{m} + \frac{\sigma_{0,1}^2}{n}\end{aligned}$$

In fact we could have a more precise formula by accounting for the second term in the H-decomposition. Lee (1990) (Theorem 2 page 38, formula 2) give the general formal for the variance that becomes in the case of a two sample U statistic of degree 1:

$$\begin{aligned}\mathbb{V}ar\left[\hat{U}\right] &= \frac{\sigma_{1,0}^2}{m} + \frac{\sigma_{0,1}^2}{n} + \frac{\sigma_{1,1}^2 - \sigma_{0,1}^2 - \sigma_{1,0}^2}{nm} \\ &= \frac{1}{nm}\left((n-1)\sigma_{1,0}^2 + (m-1)\sigma_{0,1}^2 + \sigma_{1,1}^2\right)\end{aligned}$$

where $\sigma_{1,1}^2 = \mathbb{P}\left[x < y\right]\left(1 - \mathbb{P}\left[x < y\right]\right)$. Indeed the second term of the H-decomposition would be the projection of $\mathbb{1}_{X \leq Y}$ on X, Y where we substract components of the Hájek projection to get the orthogonality between $H_{\hat{U}}^{(1)}$ and $H_{\hat{U}}^{(2)}$:

$$\begin{aligned}H_{\hat{U}}^{(2)} &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{E}\left[\mathbb{1}_{x_i \leq y_j}\middle| x_i, y_j\right] - \mathbb{E}\left[\mathbb{1}_{x_i \leq y}\middle| x_i\right] - \mathbb{E}\left[\mathbb{1}_{x \leq y_j}\middle| y_j\right] \\ &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{1}_{x_i \leq y_j} - \mathbb{E}\left[\mathbb{1}_{x_i \leq y}\middle| x_i\right] - \mathbb{E}\left[\mathbb{1}_{x \leq y_j}\middle| y_j\right]\end{aligned}$$

Indeed:

$$\begin{aligned}\mathbb{V}ar\left[H_{\hat{U}}^{(2)}\right] &= \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{V}ar\left[\mathbb{1}_{x_i \leq y_j}\right] - \mathbb{V}ar\left[\mathbb{E}\left[\mathbb{1}_{x_i \leq y}\middle| x_i\right]\right] - \mathbb{V}ar\left[\mathbb{E}\left[\mathbb{1}_{x \leq y_j}\middle| y_j\right]\right] \\ &\qquad - \mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[\mathbb{1}_{x_i \leq y}\middle| x_i\right]\right] - \mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[\mathbb{1}_{x \leq y_j}\middle| y_j\right]\right] \\ &= \frac{\sigma_{1,1}^2 - \sigma_{0,1}^2 - \sigma_{1,0}^2}{nm} - \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[\mathbb{1}_{x_i \leq y}\middle| x_i\right]\right] + \mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[\mathbb{1}_{x \leq y_j}\middle| y_j\right]\right] \\ &= \frac{\sigma_{1,1}^2 - \sigma_{0,1}^2 - \sigma_{1,0}^2}{nm}\end{aligned}$$

Since:

$$\begin{aligned}\mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[\mathbb{1}_{x_i \leq y}\middle| x_i\right]\right] &= \mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[1 - \mathbb{1}_{x_i > y}\middle| x_i\right]\right] \\ &= \mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[-\mathbb{1}_{y < x_i}\middle| x_i\right]\right]\end{aligned}$$

which under the null hypothesis that X and Y have the same distribution equals $-\mathbb{C}ov\left[\mathbb{1}_{x_i \leq y_j}, \mathbb{E}\left[\mathbb{1}_{x \leq y_j}\right]\middle| y_j\right]$.