Wilcoxon Two-Sample Test

如果我們現在要 test 兩個 continuous distribution 是不是相等的，一樣我們利用觀測值的 magnitude. 不過要做這樣的 test 之前，我們必須假設這兩個 distribution 曲線非常接近(就很像我們之前要 test 2 個 normal 是一樣的假設，我們會假設$\sigma_X^2=\sigma_Y^2$, 第一個是因為好做，雖然有提過不相等的 t test, 但是它的 degree of freedom 非常的難算，另一個是如果兩個 distribution 差太多，一般人不會把它們合起來一起考慮)。現在我們將兩組 sample $X_1,X_2,\ldots,X_{n_1}$, 與 $Y_1,Y_2,\ldots,Y_{n_2}$ 合在一起，從小到大排列，將每個數值依序給它 ranks $1,2,3,\ldots,n_1+n_2$. 如果在 sample 中有兩個以上相同的數值，那我們就針對這樣的數值採用平均的 ranks. 我們令 W 是 $Y_1,Y_2,\ldots,Y_{n_2}$ rank 的總和。若 Y 的 distribution 是在 X 的右邊，那麼我們可以預期 Y 的 value 會高於 X, 那麼 W 通常算起來也會比較大。若 m_X,m_Y 分別代表 X,Y 之 distribution 的 median, 我們可以猜測 testing $H_0:m_X=m_Y$ 與 alternative hypothesis $H_1:m_X<m_Y$ 的 critical region 一定是形如 $w>c$ 這種形式。同理，若對立假設為 $H_1:m_X>m_Y$, 那其 critical region 將會形如 $w<c$ 這種形式。

若對 W 的 distribution 有興趣的人，可以嘗試繪製 $n_1=n_2=3$ 的情形，會幫助你了解這個分配，不過這裡我們不打算去關心這個 distribution, 在實務上，若 n_1 和 n_2 都大於 7, 我們可以利用 normal 去逼近(central limit theorem)，至於個數較少的話，便可以利用手算即可，不過因此我們要去算 W 的 mean 與 variance, 這裡我們建議不要整體看 mean, 而是分別考慮每個數出現的 probability，以求其 mean.

$$\mu_{\mathrm{w}}=\frac{C_{n_2-1}^{n_1+n_2-1}}{C_{n_2}^{n_1+n_2}}[1+2+\cdots+(n_1+n_2)]$$

$$=\frac{\frac{(n_1+n_2-1)!}{(n_2-1)!}}{\frac{(n_1+n_2)!}{n_2!}}\cdot\frac{(n_1+n_2)(n_1+n_2+1)}{2}$$

$$=\frac{n_2}{n_1+n_2}\cdot\frac{(n_1+n_2)(n_1+n_2+1)}{2}=\frac{n_2(n_1+n_2+1)}{2}$$

要先算 variance 之前，我們先來考慮高中我們學過相異數兩兩乘積之和的算法：

$$\left[1+2+\cdots+(n_1+n_2)\right]^2=1^2+2^2+\cdots+(n_1+n_2)^2+2\left[1\cdot2+1\cdot3+\cdots+1\cdot(n_1+n_2)+\cdots+(n_1+n_2)(n_1+n_2+1)\right]$$

令 $S=1\cdot2+1\cdot3+\cdots+1\cdot(n_1+n_2)+\cdots+(n_1+n_2)(n_1+n_2+1)$

$$\Rightarrow\left[\frac{(n_1+n_2)(n_1+n_2+1)}{2}\right]^2=\frac{(n_1+n_2)(n_1+n_2+1)(2n_1+2n_2+1)}{6}+2S$$

$$\therefore 2S = \frac{(n_1+n_2)^2(n_1+n_2+1)^2}{4} - \frac{(n_1+n_2)(n_1+n_2+1)(2n_1+2n_2+1)}{6}$$

$$= \frac{(n_1+n_2)(n_1+n_2+1)}{12}[3(n_1+n_2)(n_1+n_2+1) - (4n_1+4n_2+2)]$$

$$= \frac{(n_1+n_2)(n_1+n_2+1)}{12}\left[3(n_1+n_2)^2 - (n_1+n_2) - 2\right]$$

$$= \frac{(n_1+n_2)(n_1+n_2+1)(n_1+n_2-1)(3n_1+3n_2+2)}{12}$$

接著我們來算 variance, 利用與 mean 同樣的想法，個別算兩兩乘積所對應的 probability, 我們就可以得到 variance 的公式

$$Var(W) = E[W^2] - \mu_w^2$$

$$= \frac{n_2}{n_1+n_2}[1^2+2^2+\cdots+(n_1+n_2)^2] + \frac{C_{n_2-2}^{n_1+n_2-2}}{C_{n_2}^{n_1+n_2}} \cdot 2S - \left[\frac{n_2(n_1+n_2+1)}{2}\right]^2$$

$$= \frac{n_2}{n_1+n_2}\frac{(n_1+n_2)(n_1+n_2+1)(2n_1+2n_2+1)}{6} + \frac{\frac{(n_1+n_2-2)!}{n_1!(n_2-2)!}}{\frac{(n_1+n_2)!}{n_1!n_2!}} \cdot 2S - \frac{n_2^2(n_1+n_2+1)^2}{4}$$

$$= \frac{n_2(n_1+n_2+1)(2n_1+2n_2+1)}{6} + \frac{n_2(n_2-1)}{(n_1+n_2)(n_1+n_2-1)} \cdot \frac{(n_1+n_2)(n_1+n_2+1)(n_1+n_2-1)(3n_1+3n_2+2)}{12}$$

$$- \frac{n_2^2(n_1+n_2+1)^2}{4}$$

$$= \frac{n_2(n_1+n_2+1)(2n_1+2n_2+1)}{6} + \frac{n_2(n_2-1)(n_1+n_2+1)(3n_1+3n_2+2)}{12} - \frac{n_2^2(n_1+n_2+1)^2}{4}$$

$$= \frac{n_2(n_1+n_2+1)}{12}[(4n_1+4n_2+2) + (n_2-1)(3n_1+3n_2+2) - 3n_2(n_1+n_2+1)]$$

$$= \frac{n_2(n_1+n_2+1)(n_1+n_2-n_2)}{12} = \frac{n_1n_2(n_1+n_2+1)}{12}$$