Suppose that $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ are random samples from continuous distributions $F$ and $G$, respectively. Wilcoxon’s two-sample test statistic $W = W(X, Y)$ is defined to be $\sum_{i=n+1}^{n+m} R_i$ where $R_i$ is the rank of $Y_i$ in the combined sample.

1. Let $T = \sum_{i=1}^{m} V_i$ where $V_1, V_2, \ldots, V_m$ are numbers sampled at random without replacement from the set $\{1, 2, \ldots, N\}$.

Show that $E(Z_i) = (N + 1)/2$ and hence $E(T) = m(N + 1)/2$.

Show that

$$E(V_i^2) = \frac{2N^2 + 3N + 1}{6} \quad \text{and} \quad E(V_i V_j) = \frac{3N^2 + 5N + 2}{12}, i \neq j,$$

and hence

$$\text{Var}(T) = \sum_{i=1}^{m} \text{Var}(V_i) + \sum_{i \neq j} \text{Cov}(V_i, V_j) = \frac{m(N-m)(N+1)}{12}.$$

Deduce that under the null hypothesis that $F = G$, the expectation and variance of Wilcoxon’s two-sample test statistic are $m(n+m+1)/2$ and $nm(n+m+1)/12$, respectively.

2. Show that $W$ can be written as

$$W = U + \frac{1}{2} m(m + 1),$$

where $U$ is the number of pairs $(X_i, Y_j)$ with $X_i < Y_j$. In other words

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} I_{i,j}, \quad \text{where} \quad I_{i,j} = \begin{cases} 1 & \text{if } X_i < Y_j, \\ 0 & \text{otherwise.} \end{cases}$$

[Hint: Let $Y_1, Y_2, \ldots, Y_m$ be the order statistics for the $y$-sample. Then $U$ is the number of pairs $(X_i, Y_j)$ with $X_i < Y_j$. For fixed $j$, the number of $X_i$ with $X_i < Y_j$ is just the rank of $Y_j$ minus the number of $Y_k$ with $Y_k \leq Y_j$, that is $R_j - j$. Now sum over $j$.]
3. Continuing from Question 2 show that, under the null hypothesis that \( F = G \),
\[
E(I_{i,j}) = \frac{1}{2}; \quad E(I_{i,j}I_{i,\ell}) = \frac{1}{3}, \ j \neq \ell; \quad E(I_{i,j}I_{k,\ell}) = \frac{1}{4}, \ i \neq k, j \neq \ell.
\]
Hence obtain the mean and variance of the Wilcoxon statistic under the null hypothesis.

4. Explain why the identity \( W = U + m(m + 1)/2 \) in Questions 2, shows that the value of \( \Delta \) which minimises
\[
\left| W(X, Y - \Delta) - \frac{m(n + m + 1)}{2} \right|
\]
is given by
\[
\hat{\Delta} = \text{median}\{Y_j - X_i, 1 \leq i \leq n, 1 \leq j \leq m\}.
\]

5. (optional) The paper “Measuring the exposure of infants to tobacco smoke,” (New England Journal of Medicine, 1984, pp. 1075-1078) compared infants who had been exposed to household smoke with those that hadn’t. The data are observations of urinary concentration of cotamine, a major metabolite of nicotine. Adapt the Wilcoxon rank sum statistic, \( W \), to test whether the mean cotanine level in the population of exposed children is more than 25 units higher than that of unexposed children. Use \( W \) (and any method you like except \texttt{wilcox.test()} in R) to form a confidence interval for the difference, \( \Delta \), between the means of the two populations, stating any assumptions.

\textbf{Unexposed:} 8, 11, 12, 14, 20, 43, 111

\textbf{Exposed:} 35, 56, 83, 92, 128, 150, 176, 208

6. You have two independent samples, \( X_1, \ldots, X_n \) and \( Y_1, \ldots, Y_m \) drawn from populations with continuous distributions. Suppose the two samples are combined and the combined set of values are put in increasing order. Let \( V_r = 1 \) if the value with rank \( r \) in the combined sample is a \( Y \) and \( V_r = 0 \) if it is an \( X \), for \( r = 1, \ldots, N \), where \( N = m + n \).

Show that, if the two populations are the same then
\[
E(V_r) = \frac{m}{N}, \quad \text{Var}(V_r) = \frac{mn}{N^2}, \quad \text{Cov}(V_r, V_s) = -\frac{mn}{N^2(N - 1)}, \ r \neq s.
\]
The general linear rank statistic is defined to be \( T = \sum_{r=1}^{N} a_r V_r \), where \( (a_r) \) are given constants.

Show that the mean and variance of the general linear rank statistic are
\[
E(T) = \frac{m}{N} \sum_{r=1}^{N} a_r \quad \text{and} \quad \text{Var}(T) = \frac{mn}{N(N - 1)} \sum_{r=1}^{N} (a_r - \bar{a})^2.
\]
Show that Wilcoxon’s \( W \) is a linear rank statistic. Use the formulae above to derive its mean and variance, under the null hypothesis that the populations are the same.
7. Explain carefully what each line of the following R-code does and what is illustrated by the resulting plots.

```r
par(mfrow=c(2,2))
for(n in c(4,5,10,40)){
x <- seq(0,n*(n+1)/2)
plot(x, dsignrank(x,n), type="h", main=paste("dsignrank(x,n =",n,"")")
}
```

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution with a symmetric density centred about an unknown median $\mu$. The Wilcoxon signed rank test statistic for the null hypothesis $H_0 : \mu = \mu_0$ is

$$S_R(\mu_0) = \sum_{i=1}^{n} R_i I_i,$$

where $R_i$ is the rank of $|X_i - \mu_0|$ in the set \{|$X_1 - \mu_0$|, $\ldots$, $|X_n - \mu_0|$\} and $I_i$ is the indicator for the event $X_i > \mu_0$.

Show that, under the null hypothesis, $S_R(\mu_0)$ has the same distribution as $T_n = \sum_{r=1}^{n} r J_r$ where $J_1, \ldots, J_n$ are i.i.d. Bernoulli variables with success probability $p = 1/2$. Hence obtain the mean and variance of $S_R(\mu_0)$ under $H_0$.

Show that $T_n$ can take any integer value between 0 and $n(n+1)/2$ and the distribution of $T_n$ is symmetric about $n(n+1)/4$. Show that

$$P(T_n = k) = \frac{u_n(k)}{2^n}, \quad k = 0, \ldots, n(n+1)/2,$$

where

$$u_n(k) = u_{n-1}(k) + u_{n-1}(k-n),$$

and hence or otherwise verify that $P(T_7 \geq 26) = 0.023$.

The following observations are pH values of synovial fluid (which lubricates joints and tendons) taken from knees of a random sample of individuals suffering from arthritis (Arthritis and Rheumatism, 1971, pp. 476-477).

$$7.02, 7.34, 7.28, 7.09, 7.45, 7.40, 7.32$$

Using the exact distribution of the signed-rank statistic, compute a 95% signed-rank confidence interval for the population median, $\mu$. Verify your answer using R.