B. Applied Statistics II

4. Consider the data in Table 1 taken from Canadian records of pure-bred dairy cattle. They give average butterfat percentages for random samples of 10 mature cows.

<table>
<thead>
<tr>
<th>Cattle type</th>
<th>Sample 1</th>
<th>Sample 2</th>
<th>Sample 3</th>
<th>Sample 4</th>
<th>Sample 5</th>
<th>Sample 6</th>
<th>Sample 7</th>
<th>Sample 8</th>
<th>Sample 9</th>
<th>Sample 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canadian</td>
<td>3.92</td>
<td>4.95</td>
<td>4.47</td>
<td>4.28</td>
<td>4.07</td>
<td>4.10</td>
<td>4.38</td>
<td>3.98</td>
<td>4.46</td>
<td>5.05</td>
</tr>
<tr>
<td>Guernsey</td>
<td>4.54</td>
<td>5.18</td>
<td>5.75</td>
<td>5.04</td>
<td>4.64</td>
<td>4.79</td>
<td>4.72</td>
<td>3.88</td>
<td>5.28</td>
<td>4.66</td>
</tr>
</tbody>
</table>

Table 1: Butter fat % for two different cattle types, 5 years and older (Sokal and Rohlf, 1981).

(a) [6 marks] State the formula for the two sample Wilcoxon test statistic $W$ and the assumptions on the samples $X = (X_1, \ldots, X_n)$ and $Y = (Y_1, \ldots, Y_m)$.

Calculate the value of $W$ for the data provided.

(b) [6 marks] Consider the null hypothesis that the distribution of average butterfat is the same for Canadian and Guernsey cattle.

(i) Using the normal approximation to the distribution of $W$ under the null hypothesis, or otherwise, test the null hypothesis at the 5% level.

[Note that $\text{Var } W = nm(n + m + 1)/12$ under the null hypothesis.]

(ii) The Wilcoxon two sample test is invariant under a large class of transformations of the data. What is this class? Explain why the test statistic is invariant.

(iii) Describe one additional method to calculate the p-value of the Wilcoxon two sample test.

(c) [5 marks] Consider $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_1$ and $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} F_2$. We assume that $F_1(t) = F_2(t + \Delta)$. State the Hodges-Lehman estimator for difference in location. How many data items of $X$ need to be corrupted for the location estimate to take arbitrarily large values?

(d) [5 marks] Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_1$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} F_2$, independent of each other. Fix a threshold $t \in \mathbb{R}$ and let $U$ and $V$ denote the number of the X’s and Y’s respectively that are less than or equal to $t$. Then $U$ and $V$ have Binomial distributions with parameters $\mathbb{P}(X \leq t)$ and $\mathbb{P}(Y \leq t)$, respectively. Consider the null-hypothesis that $F_1 = F_2$. Let

$$S = U - V$$

with null distribution

$$\mathbb{P}(S = i) = \sum_{j,k:j-k=i} \binom{n}{j} \binom{n}{k} p^{j+k}(1-p)^{2n-i-k}$$

where the unknown $p$ can be replaced by the estimate $\hat{p} = \frac{U+V}{2n}$.

(i) Give an example of $F_1 \neq F_2$ and $t$ for which the power of the test based on $S$ does not increase to 1 as $n \rightarrow \infty$.

(ii) Give an additional disadvantage of this test compared to the Wilcoxon test.
5. (a) [15 marks] (Bootstrapping) Let \( Y \) be a Poisson distributed random variable \( Y \sim \text{Po} (\lambda) \). We would like to estimate \( \theta = \text{median} (Y) \) on the basis of \( Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \text{Po} (\lambda) \).

(i) Describe two estimators for \( \hat{\theta} \). An exact formula is not required.

(ii) The aim is to estimate the standard error \( \text{se}(\hat{\theta}) \). Describe in words, or using pseudocode, the parametric bootstrap estimate of \( \text{se}(\hat{\theta}) \).

(iii) Describe a method in words, or using pseudocode, to obtain a nonparametric bootstrap estimate of \( \text{se}(\hat{\theta}) \).

(iv) Describe one method to obtain a bootstrap confidence interval. State if the method yields a first order or a second order accurate confidence interval. Explain what is meant by first order and second order accuracy.

(b) [7 marks] (Local linear regression) Consider the one-dimensional regression problem

\[
Y_i = f(x_i) + \varepsilon_i \quad \text{for } i = 1, \ldots, n
\]

with \( x_i \in \mathbb{R} \), where \( \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \) and where \( f: \mathbb{R} \to \mathbb{R} \) is an unknown twice continuously differentiable function.

For a kernel \( K_h \), local linear regression around \( x_0 \) is given implicitly through the minimisation problem

\[
(\hat{\alpha}(x_0), \hat{\beta}(x_0)) = \arg \min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{n} K_h(x_0, x_i) (y_i - \alpha(x_0) - \beta(x_0)x_i)^2
\]

such that the regression estimate is given by \( \hat{f}(x_0) = \hat{\alpha}(x_0) + \hat{\beta}(x_0)x_0 \). The estimate takes the form

\[
\hat{f}_h(x_0) = b(x_0) \left( B^T W(x_0) B \right)^{-1} B^T W(x_0) Y
\]

where \( b(x) = (1, x) \), \( B = (b(x_1)^T, \ldots, b(x_n)^T) \) and \( W(x) \) is a diagonal matrix with entries \( K_h(x_0, x_i) \).

(i) Consider a kernel of the form \( K_h(x, y) = K\left(\frac{|x-y|}{h}\right) \) for a twice continuously differentiable function \( K: \mathbb{R} \to \mathbb{R} \) such that \( K(x) \geq 0 \ \forall x \in \mathbb{R}, \int K(x) \, dx = 1 \) and \( \int xK(x) \, dx = 0 \).

The prediction error or risk is typically defined as \( R(h) = E\{(Y - \hat{f}_h(X))^2\} \) where the expectation is with respect to random new observations \( Y \) and \( x \) chosen randomly among \( (x_1, \ldots, x_n) \). Sketch qualitatively the typical behaviour of \( R(h) \) as \( h \) varies. How does this relate to the choice of \( h \)? Explain the terms undersmooth and oversmooth.

(ii) Let \( l(x_0) = b(x_0) \left( B^T W(x_0) B \right)^{-1} B^T W(x_0) \). Prove that

\[
\sum_{i=1}^{N} l_i(x_0) = 1 \quad \text{and} \quad \sum_{i=1}^{N} (x_i - x_0) l_i(x_0) = 0.
\]

[Hint: consider \( l(x_0)B \).]