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.

Cattle type \ Sample	1	2	3	4	5	6	7	8	9	10
Canadian	3.92	4.95	4.47	4.28	4.07	4.10	4.38	3.98	4.46	5.05
Guernsey	4.54	5.18	5.75	5.04	4.64	4.79	4.72	3.88	5.28	4.66

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 $X = (X_1, \dots, X_n)$ and $Y = (Y_1, \dots, Y_m)$.

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* $\operatorname{Var} W = nm(n+m+1)/12$ *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 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F_1$ and $Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} F_2$. 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 $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} F_1$ and $Y_1, \dots, Y_n \overset{\text{i.i.d.}}{\sim} F_2$, 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 $\mathbb{P}(X \leqslant t)$ and $\mathbb{P}(Y \leqslant t)$, 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 $\widehat{p} = \frac{U+V}{2n}$.

(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 $n \to \infty$.

(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 $Y \sim \mathrm{Po}(\lambda)$. We would like to estimate $\theta = \mathrm{median}(Y)$ on the basis of $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{Po}(\lambda)$.

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

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

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

   (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

$$Y_i = f(x_i) + \varepsilon_i \qquad \text{for } i = 1, \ldots, n$$

with $x_i \in \mathbb{R}$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$ 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

$$\left(\widehat{\alpha}(x_0), \widehat{\beta}(x_0)\right) = \underset{\alpha(x_0), \beta(x_0)}{\arg\min} \sum_{i=1}^{n} K_h(x_0, x_i) \left(y_i - \alpha(x_0) - \beta(x_0) x_i\right)^2$$

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

$$\widehat{f}_h(x_0) = b(x_0) \left(B^T W(x_0) B\right)^{-1} B^T W(x_0) Y$$

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

   (i) Consider a kernel of the form $K_h(x, y) = K(\frac{y-x}{h})$ for a twice continuously differentiable function $K \colon \mathbb{R} \to \mathbb{R}$ such that $K(x) \geqslant 0 \; \forall x \in \mathbb{R}$, $\int K(x)\, dx = 1$ and $\int x K(x)\, dx = 0$.

   The prediction error or risk is typically defined as $\mathrm{R}(h) = E\left\{(Y - \widehat{f}_h(X))^2\right\}$ 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 $\mathrm{R}(h)$ 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* $l(x_0)B$.]