1. 探討某地區五年來乳癌發生個案如何受女性荷爾蒙是否使用（1=是；0=否）之影響，相關資料如下：

年齡	女性荷爾蒙使用	乳癌個案	人年*
＜50	是	10	20,000
	否	10	60,000
≧50	是	60	40,000
	否	40	80,000
合計		120	200,000

（註*：人年＝追蹤時間x追蹤人數）

若定義隨機變數Y爲乳癌發生個數，每單位人年發生乳癌個案（即乳癌發生率）以λ表示。請回答下列問題：

【問題 1】根據上述資料，寫出所有婦女發生乳癌個案之機率密度函數。（4 分）

【問題 2】依據【問題 1】，寫出所有婦女發生乳癌個案之累積分佈函數。（4 分）

【問題 3】依據【問題 1】，寫出其對數概似函數(log- likelihood)。（4 分）

【問題 4】根據上述資料求λ之最大可能概似值(Maximun Likelihood Estimate, MLE)。（4 分）

【問題 5】求上述【問題 4】λ估計值之變異數。（4 分）

【問題 6】若吾人有興趣估計使用女性荷爾蒙者較非使用者之乳癌發生率之比值（以θ參數表示），以年齡＜50 爲例，請使用再參數化(reparameterization)描述所須估計之母體參數及虛無假說。（4 分）

【問題 7】根據上述表中資料，寫出上述【問題 6】再參數化後之對數概似函數。（4 分）

【問題 8】根據上述表中資料，利用【問題 7】求θ之最大概似值。（4 分）

【問題 9】寫出適合上述表中資料之迴歸數學模式及其模式內相關參數。（4 分）

【問題 10】根據上述表中資料及【問題 9】之迴歸模式，在控制年齡後求θ之估計值。（4 分）

見背面

2. 在一個探討心血管疾病與熱量攝取相關性研究，研究者先選取 n 位心血管疾病患者作爲個案組，然後再去尋找一位與患者之年齡、性別相配對的個體作爲對照，形成有 n 位個體之對照組，對此 $2n$ 位個體資料，調查每一位個體之每日平均熱量攝取值，進行相關分析。以 y_{ij} 表示此 $2n$ 個體之每日平均熱量攝取值，$i=1,...,n$，$j=1$（個案組）、0（對照組），二組之平均值分別爲 $\mu_1=E(y_{i1})$ 及 $\mu_0=E(y_{i0})$，二組之變異數爲 σ_1^2 及 σ_0^2，假設 $\sigma_1^2=\sigma_0^2=\sigma^2$。

【問題 1】檢定心血管疾病與熱量攝取間的相關可以檢定 $H_0:\mu_1=\mu_0$ vs. $H_1:\mu_1\neq\mu_0$ 進行，請解釋爲何這個作法是合理的。（5 分）

檢定 H_0 的檢定方法，一般是透過 $d_i=y_{i1}-y_{i0}$ 的設定，建立檢定統計量

$$t^*=\frac{\bar{d}}{\sqrt{\hat{V}_1(\bar{d})}},$$

其中，$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}d_i=\frac{1}{n}\sum_{i=1}^{n}y_{i1}-\frac{1}{n}\sum_{i=1}^{n}y_{i0}=\bar{y}_1-\bar{y}_0$ 且

$\hat{V}_1(\bar{d})=\frac{1}{n}\hat{V}(d_i)=\frac{1}{n}\left\{\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2\right\}$。

【問題 2】請先說明 t 分布的定義，繼而驗證在什麼假設之下 t^* 的分布會是 t 分布。（10 分）

另一種檢定 H_0 的統計檢定量爲：

$$t^{**}=\frac{\bar{d}}{\sqrt{\hat{V}_2(\bar{d})}},$$

其中，$\hat{V}_2(\bar{d})=\hat{V}(\bar{y}_1)+\hat{V}(\bar{y}_0)-2\widehat{Cov}(\bar{y}_1,\bar{y}_0)=\frac{2\hat{\sigma}^2}{n}-2\widehat{Cov}(\bar{y}_1,\bar{y}_0)$。

【問題 3】請說明計算 $\hat{\sigma}^2$ 和 $\widehat{Cov}(\bar{y}_1,\bar{y}_2)$ 的式子爲何？（6 分）

接次頁

（續題 2.）

以上之配對研究設計在方法學特稱爲配對個案對照（matched case control）設計，這種研究設計的特點在於藉由配對設計，消除年齡、性別等干擾因子對檢定 H_0 之影響，但它的附帶缺點是個案對照二組之間或許存在相依性而須考慮它對檢定的影響。分析二組之間相關性（註：不是疾病與攝取熱量間相關）對檢定的影響，必須根據 t^{**} 進行。將 $\hat{V}_2(\bar{d})$ 改造成如下形式：

$$\hat{V}_3(\bar{d})=\hat{V}(\bar{y}_1)+\hat{V}(\bar{y}_0)-2\rho\sqrt{\hat{V}(\bar{y}_1)\hat{V}(\bar{y}_0)}=\frac{2\hat{\sigma}^2}{n}-2\rho(\frac{\hat{\sigma}^2}{n})=\frac{2(1-\rho)\hat{\sigma}^2}{n}\ ,$$

其中 ρ 是描述每一配對內的二個體間的相關係數。在實際分析時可以由另一個資料得到 ρ 的估計值，然後將之代入 $\hat{V}_3(\bar{d})$，得到以下檢定統計量數值

$$t^{***}=\frac{\bar{d}}{\sqrt{\hat{V}_3(\bar{d})}}$$

進行 H_0 檢定。

【問題 4】討論建立 t^{*}、t^{**} 和 t^{***} 三種檢定方法的統計想法。（9 分）

3. 欲探討長期飲用香椿茶是否具改善血糖之效果，研究者將 n 位高血糖者隨機分派成 2 組，其中有 n_1 位高血糖者被分派至持續飲用香椿茶 6 個月的「實驗組」，其餘 $n-n_1$ 位則爲不飲用香椿茶的「控制組」。令 Y_{i0} 爲第 i 位高血糖者在剛進入此研究時所測之飯前血糖值，Y_{i1} 爲第 i 位高血糖者在 6 個月後所測之飯前血糖值；X_i 爲代表第 i 位高血糖者被分派至實驗組 $(X_i=1)$ 或控制組 $(X_i=0)$，即 $X_1=X_2=...=X_{n_1}=1$ 且 $X_{n_1+1}=X_{n_1+2}=...=X_n=0$，$i=1,...,n$。假設在給定 X_i（分派組別）的條件下，(Y_{i0},Y_{i1}) 之條件聯立分布爲一二元常態分布，即可表示爲 $(Y_{i0},Y_{i1})\,|\,X_i$~Bivariate Normal$(\mu_{0X_i},\mu_{1X_i},\sigma_0^2,\sigma_1^2,\rho)$。

【問題 1】若研究者僅提供實驗組降低血糖的人數爲 d_0，控制組降低血糖的人數爲 d_1 的資料，根據所提供資料與上述機率模式假設，請寫出 (d_0,d_1) 與 $\{(Y_{i0},Y_{i1}),i=1,...,n\}$ 之數學關係、d_0 與 d_1 的機率分布、以及此研究欲估計或檢定的參數。（10 分）

見背面

（續題 3.）

【問題 2】若研究者提供 $\{(Y_{i0}, Y_{i1}), i = 1, ..., n\}$ 的觀察值，根據所提供資料與上述機率模式假設，請寫出此研究欲估計或檢定的參數、此參數的估計量、以及此估計量的機率分布。（10 分）

【問題 3】請寫出【問題 1】與【問題 2】的檢定統計量，並比較之。（10 分）