# COMP0118: Coursework 2

Mingzhou Hu

February 2019

## Part 1

### Q1

(a) I set the random seed to three for result reproducibility. The mean and standard deviation for sample1 are 0.9663 and 0.1688 respectively and these for sample2 are 1.5028 and 0.1996 respectively. From the results above, we can know that the values are as expected.
(b) The two-sample t-statistic of the two samples is -10.2616, with h=1 and p-value=$1.1803 \times 10^{-13}$. We can know that from the values above we reject the null hypothesis and the two samples are generated from distributions with different means.
(c) **i.** The design matrix has two columns and 50 rows, with one for the first column of the first 25 rows and the second column of the last 25 rows, and zero for the first column of the last 25 rows and the second column of the first 25 rows. $dim(X)$ is 2 because $X$ is made of two column linear independent vectors.

**ii.** $Y = X\beta \rightarrow X^TY = X^TX\beta \rightarrow (X^TX)^{-1}X^TY = \beta \rightarrow X(X^TX)^{-1}X^TY = X\beta$. Since $PxY = X\beta$, we can deduce that $Px = X(X^TX)^{-1}X^T$. Firstly, we check the idempotence($PxPx = Px$) of $Px$: $PxPx = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T$, since $(X^TX)^{-1}X^TX = I$, we can get that $PxPx = X(X^TX)^{-1}X^T = Px$. Then we check the symmetry($Px = Px^T$) of $Px$: $Px^T = (X(X^TX)^{-1}X^T)^T = (X^T)^T((X^TX)^{-1})^TX^T = X((X^T)^TX^T)^{-1}X^T = X(XX^T)^{-1}X^T = X(X^TX)^{-1}X^T = Px$. $Px$ for $C(X)$ in this question is a $50 \times 50$ matrix, with 0.04 for the first 25 columns of the first 25 rows and the last 25 columns of the last 25 rows, and zero for the first 25 columns of the last 25 rows and the last 25 columns of the first 25 rows. The trace of $Px$ is 2, which means the trace of the projection matrix is the dimension the column space.

**iii.** $\hat{Y} = PxY = X(X^TX)^{-1}X^TY$, $\hat{Y}$ is one column vector with 50 rows whose first 25 elements are 0.9663(the mean of sample1) and last 25 elements are 1.5028(the mean of sample2). $\hat{Y}$ means that the fitted value for each group. Assume that there is a linear model $Y = X\beta$, $\beta$ is unknown, Y can be found in the column space of $X : C(X)$. So $C(x)$ is the estimation space.

**iv.** $Rx$ is a $50 \times 50$ matrix, with -0.04 for the first 25 columns of the first 25 rows and the last 25 columns of the last 25 rows, and zero for the first 25 columns of the last 25 rows and the last 25 columns of the first 25

rows, but the values for $Rx(1,1), Rx(2,2), ..., Rx(50.50)$ are 0.96. Firstly, we check the idempotence($RxRx = Rx$) of $Rx$: $RxRx = (I - Px)(I - Px) = II - 2IPx + PxPx = I - 2Px + Px = I - Px = Rx$. Then we check the symmetry($Rx = Rx^T$) of $Rx$: $Rx^T = (I - Px)^T = I^T - Px^T = I - Px = Rx$. Rx satisfies the key properties of a perpendicular projection operator, so it is also a perpendicular projection operator.

**v.** $\hat{e} = Y - \hat{Y} = (I - Px)Y = RxY$, $\hat{e}$ is one column vector with 50 rows, which is stored at ehat(MATLAB). Therefore the dimension of $C(X)^{\perp}$ is one.

**vi.** The angle between $\hat{e}$ and $\hat{Y}$ can be computed by this equation: $\theta = arccos(\hat{e} \cdot \hat{Y})$ and we get the result $\theta = \frac{\pi}{2}$. From the result, we can know that $\hat{e}$ is perpendicular to $\hat{Y}$, therefore the result is expected.

**vii.** Assume that we have a linear model: $Y = X\beta + e$, and we need to minimize the sum of square errors($SSE = e^T e$) and get a minimising $\hat{\beta}$. Therefore, the general formula is known as a least squares estimate. From Q1(c(ii)), we can know that $\hat{\beta} = (X^T X)^{-1} X^T Y$. We obtain that $\hat{\beta}(1)$ is 0.9663 and $\hat{\beta}(2)$ is 1.5028, which are the means for sample1 and sample2 respectively.

**viii.** In this case, n=50 and $dim(X)$=2, and we obtain that $\hat{\sigma}^2 = 0.0342$. $\hat{\sigma}^2$ is an unbiased estimate of the error in the prediction, the upper term $e^T e$ is the sum of square errors which is divided by the number of degree of freedom. So $\hat{\sigma}^2$ is also known as the mean squared error.

**ix.** We compute that $S_{\hat{\beta}} = \hat{\sigma}^2 (X^T X)^{-1} = \begin{bmatrix} 0.0014 & 0 \\ 0 & 0.0014 \end{bmatrix}$, the covariance matrix has only zero off-diagonal terms. Therefore, the model parameters are independent from each other. The standard deviation can be computed by $\sqrt{S_{\hat{\beta}}(1,1)} = 0.0370$.

**x.** A contrast vector is a vector whose elements sum to zero. We derive the contrast vector $\lambda = [1 \ -1]^T$ to compare the group differences in the means. Then $\lambda^T \beta = \lambda_1 \beta_1 + \lambda_2 \beta_2 = \beta_1 - \beta_2 = 0$, and we set the $\beta_1 = \beta_2 = \beta_0$. Substituting this into our GLM model, we can get $Y = (X_1 + X_2)\beta_0 + e = X_0 \beta_0 + e$. Therefore, $X_0 = X_1 + X_2$, which is one column vector with 1 for all rows.

**xi.** Firstly, we need to calculate the new $\hat{\beta_0}$ by the equation: $\hat{\beta_0} = (X_0^T X_0)^{-1} X_0^T Y = 1.2346$ , and then calculate the new $\hat{e_0}$: $\hat{e_0} = Y - X_0 \hat{\beta_0}$. The additional error=$\hat{e_0} - \hat{e} = 0.2682$. The formula to estimate F-statistic is: $F = \frac{\frac{SSR(X_0) - SSR(X)}{v1}}{\frac{SSR(X)}{v2}}$, $SSR(X_0) = \hat{e_0}^T \hat{e_0}$, $SSR(X) = \hat{e}^T \hat{e}$, $v1 = tr(P_X - P_{X_0}) = 1$, $v2 = tr(I - P_X) = 48$. We computed that $F = 105.2998$, and the degrees of freedom are 1 and 48.

**xii.** The t-statistic is -10.2616 which is exactly the same as that calculated in (b), so it is identical to that in (b).

**xiii.** The parameters $\hat{\beta}_1$ and $\hat{\beta}_2$ mean that the means for group1 and group2 respectively. Their ground truth values should be $\hat{\beta}_{1gt} = 1$ and $\hat{\beta}_{2gt} = 1.5$ respectively.

**xiv.** The projection of the ground truth deviation $e$ into $C(X)$ is : $e = y - X\hat{\beta}$. We obtain that $e$ is one column vectors with 50 rows and it is stored at egtd in MATLAB. $e$ represents the difference between $\hat{\beta}_{gt}$ and $\hat{\beta}$.

**xv.** The projection of the ground truth deviation $e_{gtd}$ into $C(X)_{\perp}$ is:

$e_{gtd} = (I - P_X)e = (I - P_X)(Y - X\beta) = (I - P_X)Y - (I - P_X)X\beta$, since $(I - P_X)Y = \hat{e}$ and $I - P_X$ and $X\beta$ are orthogonal, $e_{gtd} = \hat{e}$. The result means that the projection of the ground truth deviation $e_{gtd}$ into $C(X)_\perp$ is the same as the measured error $\hat{e}$.

(d) **i.** The design matrix are three columns with 50 rows. All entries are one in the first column. The second and third columns are the same as the design matrix in ci above. $dim(X)$ is still 2 because $X$ only have two column linear independent vectors(the second and third column).

**ii.** We use the operator pinv instead of inv to compute the pseudoinverse of $X^T X$ in MATLAB. The result of $P_X$ we obtained is the same as that in (c) and the estimation space is also the same as that in (c).

**iii.** The appropriate contrast vector $\lambda = [0\ 1\ -1]^T$. Then $\lambda^T\beta = \lambda_0\beta_0 + \lambda_1\beta_1 + \lambda_2\beta_2 = \beta_1 - \beta_2 = 0 \rightarrow \beta_1 = \beta_2$. Substituting this into our GLM model, we can get $Y = X_0\beta_0 + (X_1 + X_2)\beta_1 + e = X_0\beta_0 + (X_3)\beta_1 + e$. Both $X_0$ and $X_3$ are one column vectors with 50 rows and all entries are one, so $X_0$=$X_3$. Hence, we can get $Y = X_0(\beta_0 + \beta_1) + e = X_0(\beta_*) + e$. $X_3$ is the reduced model which is the same as that in (c).

**iv.** We obtain that t=10.2616, which is the same as that in (b) and (c).

**xv.** Parameter $\hat{\beta_0}$ means a bias term applied to both groups. $\hat{\beta_1}$ and $\hat{\beta_2}$ are the differences between the means for group1 and group2 and $\hat{\beta_0}$ respectively.

(e) **i.** The design matrix are two columns with 50 rows. All the entries in the first column are one, the first 25 rows of the second column are one and the last 25 rows of the second column are zero. The dimension of its column space $dim(X)$ is 2(2 column linear independent vectors).

**ii.** The appropriate contrast vector $\lambda = [0\ 1]^T$. Then we have $\lambda^T\beta = \lambda_0\beta_0 + \lambda_0\beta_0 = \beta_1 = 0$. Hence, we can get $Y = X_0\beta_0 + e$, the same reduced model as that in (c) and (d).

**iii.** We obtain that t=10.2616, which is the same as that in (b), (c) and (d).

**iv.** Parameter $\hat{\beta_0}$ means the mean for group2 and $\hat{\beta_1}$ is the difference between the means for two groups.

(f) We cannot test the same hypothesis with an simpler model: $Y = X_0\beta_0 + e$, because this model already assumes that all samples have the same mean. In order to test that the two groups with different means we need a more complex model.

**Q2**

(a) The t-statistic of is -10.2881, which is very closer to that in Q1, and h=1 and p-value=$2.8143 \times 10^{-10}$. We can know that from the values above we reject the null hypothesis.

(b) **i.** The design matrix X has 27 columns and 50 rows, which is $\begin{bmatrix} 1...1 & 1...1 \\ 1...1 & 0...0 \\ I & I \end{bmatrix}^T$, I is an identical matrix with dimensions $25 \times 25$. The rank of $X$ is 26, because $X_0 = X_2 + X_3 + ... + X_{26} + X_{27}$(26 linear independent vectors).

**ii.** The appropriate contrast vector is $\lambda = [0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]^T$

and the reduced model is $\begin{bmatrix} 1...1 & 1...1 \\ I & I \end{bmatrix}^T$.

**iii.** We get that t=-10.2881, which is identical to that in (a).

# Part 2

## Q1

(a) The random seed is set to ten for result reproducibility. t-statistic is -3.3310, which is appropriate but differient from part 1. The p-value is 0.0060 in this case.

(b) **i.** One-dimensional array $D$ stores 14 observations with the first 6 entries for group 1 and the rest for group 2.

**ii.** The valid permutations of D for group1 can be computed by the function: combnk(D, 6), 6 is the size of group1. Then loop over length(C1) and use the function setdiff to compute the valid permutations of D for group2 in case of repetition.

**iii.** The t-statistics between permutations for group1 and group2 are stored in t-statistic array. The figure(Figure 1) below shows the empirical distribution of the t-statistic with a histogram. As expected, the distribution is Gaussian distribution.

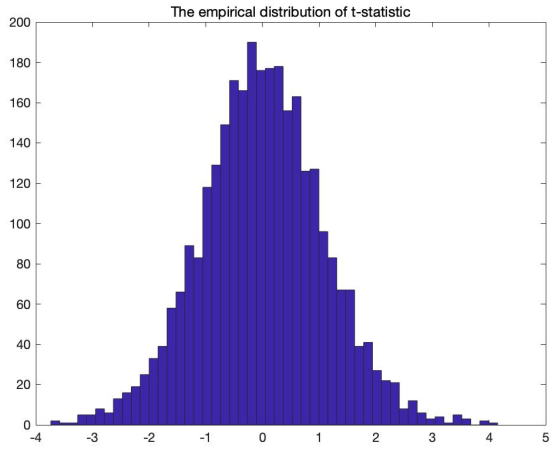


Figure 1: The empirical distribution of t-statistic

**iv.** We get that p-value is 0.0013, which is lower than that in (a), but it is appropriate.

(c) We obtain that p-value is 0.0013 using the difference between the means as the test statistic, which is same as that in (b) but lower than that in (a). The figure(Figure 2) below shows the empirical distribution of the difference between

means with a histogram. As expected, the distribution is Gaussian distribution.

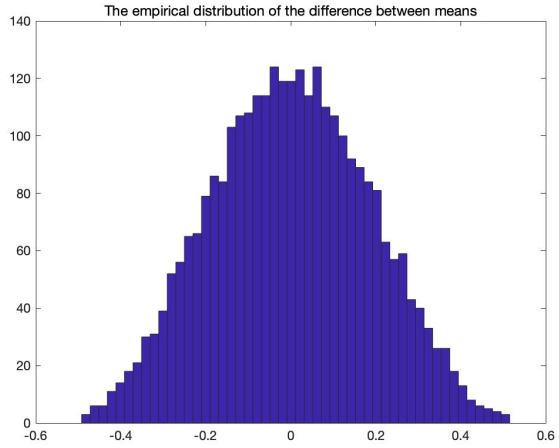


Figure 2: The empirical distribution of the difference between means.

(d) **i.** We use function randperm to generate a random set of permutations and we obtain that p-value is 0.001 with t-statistic and 1000 permutations.

**ii.** The p-value(0.001) obtained in this case is lower than that of (b)(0.0013) and (c)(0.0013).

**iii.** We use sort function to sort the elements of each row of permutation and use function unique to get the position of elements without repetition. and then we can compute the number of duplicates which is 151(1000-849(id1 in MATLAB)). The new p-value without duplicates we obtain is 0.002, which is higher than that in (di) with duplicates.

## Q2

(a) We choose the GLM model in part 1-1(c): $Y = X_1\beta_1 + X_1\beta_1 + e$. The design matrix has 2 columns and 16 rows, with one for the first column of the first 8 rows and the second column of the last 8 rows, and zero for the first column of the last 8 rows and the second column of the first 8 rows. $dim(X)$ is 2. because $X$ is made of two column linear independent vectors. The t-statistic for each voxel can be computed by $t = \frac{\lambda^T\hat{\beta}}{\sqrt{\lambda^T S_{\hat{\beta}}\lambda}}$, $\lambda = [1 \ -1]^T, \hat{\beta} = (X^TX)^{-1}X^TY$ and $S_{\hat{\beta}} = \hat{\sigma}^2(X^TX)^{-1}$. We only keep the t-statistic for ROI and we obtain that $t_{max} = 6.5294$.

(b) We use the same strategy as in 1-(b), such as the functions combnk and setdiff. The figure(Figure 3) below shows the empirical distribution of the max t-statistic with a histogram. As expected, the distribution is Gaussian distribution.

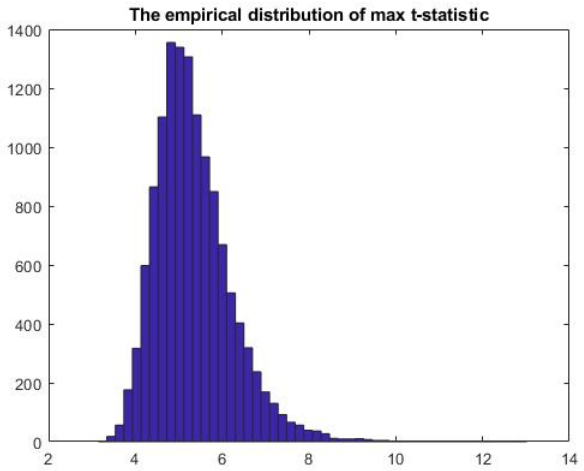


Figure 3: The empirical distribution of the max t-statistic

(c) The multiple-comparisons-corrected p-value is 0.0918 by finding the percentage of the permutations with maximum t-statistic greater than that of the original labeling.
(d) The maximum t-statistic threshold corresponding to p-value of 5% is 6.9383, which computed by function prctile.