COMP0118-Coursework2

Francesco Seracini

March 2025

1 General Linear Model (GLM) and Permutation Testing

1.1 General Linear Model

1.1.a

In this section I simulated sampled data from two groups with the same unknown variance and different means, respectively 1.5 e 2, considering the stochastic component as a normal distribution with mean equal to zero and standard deviation equal to 0.2. I computed the mean and standard deviation for the two samples and I obtained: $mean\_sample1 = 1.4674$, $std\_sample1 = 0.215\,00$, $mean\_sample2 = 2.0395$, $std\_sample2 = 0.186\,52$. The results are expected because we anticipated the means to be close to 1.5 and 2.0, but not exactly equal due to random variability, and the standard deviations to be approximately 0.2.

1.1.b

I used Matlab's built-in function *ttest2* to compute the two-sample t-statistic and I obtained $h = 1$, which means I reject the null hypothesis, that the data come from independent random samples from normal distributions with equal means, at the 5% significance level. The actual value of the t-statistic is -8.9900 that reflects our understanding that the second group has a larger mean than the first one.

1.1.c

Here I computed the t-statistic with the following GLM model: $Y = X_1\beta_1 + X_2\beta_2 + e$. The design matrix of this model can be seen in *Table 1*, where$X_{ij} = 1$ if observation Y_i belongs to group j, otherwise 0; its columnar space has dimension: $dim(C(X)) = rank(X) = 2$, which implies that there are two degrees of freedom in the estimation space. To derive the general formula of the perpendicular projection operator $P(X)$ we have followed the steps in *Table 2*. We can demonstrate that it respects the key properties of a perpendicular projection operator with the operations in *Table 3*. I used the previous formula to determine the $P(X)$ for the $C(X)$ in question and I computed its trace that is again equal to 2 because $tr(P_X) = rank(X) = dim(C(X)) = 2$ and it represents the number of independent parameters estimated by the model. The projection of Y onto $C(X)$ is: $\hat{Y} = P_X Y$ and it represents the estimated response values based on the model. It belongs to *C(X)* space, which is also called the estimation space because it contains all possible predictions that the model can make. The matrix $R_X = I - P_X$ projects onto the orthogonal complement of $C(X)$, the error space $C(X)\perp$, and we can simply verify that R_X is also a perpendicular projection operator (*Table 4*) but onto the residual space. As I used before P_X to determine the projection of Y into $C(X)$, I used R_X to determine $\hat{e}$, the projection of Y into $C(X)\perp$. It represents the estimated errors or residuals, capturing the deviation of Y from its predicted values. We can calculate the dimension of $C(X)\perp$ by using the formula: $dim(C(X)) = tr(I - P_x) = n - 2 = 38$. I proceeded to calculate the angle between $\hat{e}$ and $\hat{Y}$ using the formula: $\cos\theta = \frac{\hat{e}^T\hat{Y}}{\|\hat{e}\|\|\hat{Y}\|}$, obtaining $90°$, as expected. In fact, since $\hat{e}$ belongs to $C(X)\perp$ and $\hat{Y}$ belongs to $C(X)$, the two vectors are orthogonal and: $\hat{e}^T Y = 0$. I derived the general formula for estimating the model parameters of any GLM: $\hat{\beta} = (X^T X)^{-1} X^T Y$ (*Table 5*). It minimizes the sum of squared residuals, which is why it is called the least squares estimate. I used this formula for the actual data used in this project and I obtained: $\hat{\beta} = \begin{bmatrix} 1.4674 \\ 2.0395 \end{bmatrix}$. I proceeded to estimate the variance of the stochastic component $\hat{e}$ with the formula: $\hat{\sigma}^2 = \frac{\hat{e}^T\hat{e}}{n - \dim(X)}$, obtaining $\sigma^2 = 4.0508 \times 10^{-2}$. This value is called the mean squared errors (*MSE*) because it represents the average squared deviation of the observed values from their predicted values. Using this value and the covariance matrix of the estimated model parameters, calculated as: $S_{\hat{\beta}} = \hat{\sigma}^2 (X'X)^{-1}$, I estimated the standard deviation of the model parameters $\hat{\beta}$ (std$(\hat{\beta}_i) = \sqrt{(S_{\hat{\beta}})_{ii}}$), obtaining std$(\hat{\beta}) = \begin{bmatrix} 4.5004 \times 10^{-2} \\ 4.5004 \times 10^{-2} \end{bmatrix}$. For testing the hypothesis that the means of the two groups are equal, I set up the contrast vector $\lambda = \begin{bmatrix} 1 & -1 \end{bmatrix}^T$ and I computed the consequent reduced model X_0, obtained by enforcing $\lambda^T\beta = 0$, as a matrix with only one column of ones. I then calculated the additional error due to the constraint, by replicating the previous passages, but using the design matrix X_0 of the reduced model instead of the original X. I called ν_1 and ν_2, respectively the degrees of freedom of the numerator and the denominator of the F-statistic, calculated as: $\nu_1 = \text{tr}(P_X - P_{X_0}) = 1$, $\nu_2 = \text{tr}(I - P_{X_0}) = 38$, where ν_1 is also the number of linear restrictions. At this point, the F-statistic of comparing the reduced model to the full one, obtained by: $F_{\nu_1,\nu_2} = \frac{(SSR(X_0) - SSR(X))/\nu_1}{SSR(X)/\nu_2}$, where $SSR(X_0)$ and $SSR(X)$ are the squared errors of the two models, gives as a result of $F_{\text{stat}} = 8.0819 \times 10^1$. This high value confirms with great certainty that the two sample groups have a different mean. With the data calculated so far I also computed the t-statistic, as: $t = \frac{\lambda'\hat{\beta}}{\sqrt{\lambda' S_{\hat{\beta}} \lambda}}$, obtaining -8.9900, which is identical to the result obtained in $1.1.b$. In the GLM model used in this section

$\hat{\beta}1 = 1.4674$ and $\hat{\beta}2 = 2.0395$ represent respectively the mean of the first group and the mean of the second group. Their ground truth values should be the means assigned in $1.1.a$, 1.5 and 2.0, that are slightly different from the estimated values I obtained due to the presence of the stochastic error e. I computed the projection of the ground truth deviation e, the noise simulated in $1.1.a$, in $C(X)$ using the operator P_X obtaining an array with the first 20 elements, the ones relating to the first group, equal to -3.2644×10^{-2} and the other 20 elements equal to 3.9530×10^{-2}. We can notice that this values are exactly the difference between $\hat{\beta}$ and the ground truth β. This result is expected considering the relation: $P_X e = X(\hat{\beta} - \beta)$ (*Table 6*). Moreover, computing the projection of the ground truth e into $C(X) \perp$ we can notice that the result is exactly the same as $\hat{e}$ and this can also be demonstrated algebraically as shown in *Table 7*.

1.1.d

In this section I used the GLM model: $Y = X_0\beta_0 + X_1\beta_1 + X_2\beta_2 + e$, where X_0, called the intercept, has all its entries equal to 1 and it's used to describe the common effect between the groups. Obviously the dimension of the columnar space of this design matrix is still 2, because X_0 is not linearly independent from X_1 and X_2. For this reason I used the function $pinv$ to calculate P_X, obtaining a result slightly different than the one found in in $1.1.c$. In fact, even though in theory both models project onto the same subspace (the 2-dimensional space spanned by X_1 and X_2), using $pinv$ in the second model alter the projection matrix P_X due to how the pseudo-inverse handles the redundancy of the columns. The goal is to test whether the two groups have different means, which translates to testing the hypothesis:$H_0 : \beta_1 = \beta_2$, or $H_0 : \beta_1 - \beta_2 = 0$, so the contrast vector has to extract the difference between the two groups means and it becomes: $\lambda = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}^T$. The reduced model X_{red} is a matrix that has as first column X_0 and as second $X_1 + X_2$. I computed the t-statistic using the same formula of $1.1.c$ and I obtained -8.9900, that is the same result found in $1.1.a$ and in $1.1.c$. In this model the parameter β_0 represents the baseline mean level of Y when both $X_1 = 0$ and $X_2 = 0$, meaning when neither group effect is present. However, since every observation belongs to either group 1 or group 2, this situation never actually occurs, making β_0 more of an abstract reference level. The parameter β_1 captures the deviation from β_0 when an observation belongs to group 1 ($X_1 = 1$), indicating how much the mean response differs for this group. Similarly, β_2 represents the deviation from β_0 when an observation is in group 2 ($X_2 = 1$), quantifying the difference in mean response for that group.

1.1.e

In this section, I used the GLM model: $Y = X_0\beta_0 + X_1\beta_1 + e$. Compared to the previous model, we have removed X_2, but the dimension of the column space of the design matrix remains the same since X_2 was linearly dependent on X_0 and X_1. Therefore, the dimension of the column space of X, denoted as $\dim(C(X))$, is still 2. This means that the estimation space is unchanged, as X_0 and X_1 span the same subspace as in the previous case. To test whether the two groups have different means, we consider the hypothesis $H_0 : \beta_1 = 0$. This is equivalent to checking if the mean response for group 1 differs from the baseline effect. The contrast vector must isolate β_1, which leads to $\lambda' = \begin{bmatrix} 0 & 1 \end{bmatrix}^T$. The reduced model X_{red} is obtained by imposing $\beta_1 = 0$, which simplifies the model to $Y = X_0\beta_0 + e$. This means that under the null hypothesis, there is only a single overall mean, without a distinction between groups. I computed the t-statistic using the same formula as in the previous cases and obtained the same result. This confirms that removing X_2 did not alter the statistical inference. The reason is that the estimation space remains the same: in the previous model, X_2 was linearly dependent on X_0 and X_1, meaning that removing it does not affect the projection space. Regarding the interpretation of the model parameters, β_0 represents the overall mean level of Y when $X_1 = 0$, for observations in group 2. The parameter β_1 represents the deviation from β_0 for group 1, quantifying the difference in mean response between the two groups.

1.1.f

The model: $Y = X_0\beta_0 + e$ does not allow for direct hypothesis testing of $H_0 : \beta_1 = \beta_2$, because it does not estimate separate group means.

1.2 Paired t-test for comparing the means from repeat measurements from the same group of subjects

1.2.a

In this section, I treated the simulated data from $1.1.a$ as repeated observations from the same group of subjects. First, I used Matlab's built-in function $ttest$ to compute the paired t-statistic, obtaining -9.2097. This result is consistent with the previous test but has an even greater absolute value than the t-statistic estimated with the two-sample t-test. That is because the paired t-test eliminates variation between subjects, considering only within-subject differences. As a result, the standard deviation of differences s_D, that is at the denominator of the formula used in $ttest$, is typically smaller than the pooled standard deviation from the independent groups and so its absolute value increases.

1.2.b

In this section, I computed the paired t-statistic using the GLM model: $Y = X_0\beta_0 + X_1\beta_1 + \sum_{i=1}^{N} S_i s_i + e$, where X_0 is the intercept, X_1 is an explanatory variable indicating different time points, and S_i are dummy variables representing individual subjects. The parameters s_i account for subject-specific effects. Since there are N subjects, X has $N + 2$ columns. However, the subject dummies introduce linear dependence, making rank$(X) = N + 1 = 21$. To test the hypothesis that the means at different time

points are equal, I used the contrast vector $\lambda' = [0 \quad 1 \quad 0 \quad \dots \quad 0]^T$, which isolates β_1, representing the difference in means between the two time points. Finally, I computed the t-statistic and obtained a result with the same absolute value but opposite sign compared to the one from the paired t-test. This difference in sign arises because the paired t-test computes differences as $Y_1 - Y_2$, while in the GLM model X_1 is being defined as 1 for the second time point and 0 for the first, effectively computing the contrast in the opposite direction as $Y_2 - Y_1$.

2 Permutation Testing

2.1

2.1.a

Here, I simulated a similar dataset as in section 1.1 with sizes of $n_1 = 6$ and $n_2 = 8$ and I used the $ttest2$ function to compute the t-statistic obtaining: $t - statistic = -3.7955$, $p - value = 2.5504 \times 10^{-3}$.

2.1.b

In this section, I computed the exact permutation-based p-value for testing whether the two groups have different means. First, I constructed a one-dimensional array D containing all observations. Then, I generated all valid permutations of D, while preserving the sample sizes of each group, using MATLAB's built-in function $nchoosek$. Then, I calculated the t-statistic using $ttest2$ for each permutation, constructing the empirical distribution of the t-statistic. Finally, I determined the p-value by finding the proportion of permuted t-statistics with an absolute value greater than or equal to the one of the original labeling. The p-value found is $p_{\text{perm}} = 3.6630 \times 10^{-3}$, which is slightly larger than the p-value obtained from the original labeling. This difference is probably due to the fact that permutation test does not assume normality like the t-test but instead builds a distribution directly from the data. The histogram of the empirical distribution of t-statistics (*Figure 1*), confirming that the observed t-statistic lies in the extreme tail of the distribution, supports the rejection of the null hypothesis.

2.1.c

In this section, I repeated the permutation test using the difference in means as the test statistic instead of the t-statistic. I first computed it for the orginal samples, obtaining -0.39086 . Then, I generated all permutations, while maintaining group sizes, and I calculated the mean difference for each one of them. The p-value obtained is $p_{\text{perm}} = 3.6630 \times 10^{-3}$, identical to the one obtained in $2.1.b$. This confirms that the choice of test statistic does not significantly impact the permutation test results. The histogram in *Figure 2* shows the empirical distribution of permuted mean differences. As before, the observed statistic falls in the far end of the distribution, telling us that we have to reject the null hypothesis.

2.1.d

In this section, I estimated the p-value using only 1000 randomly generated permutations, including always the original labeling, instead of computing all possible ones. The resulting approximate p-value is $p_{\text{perm}} = 5.0000 \times 10^{-3}$, slightly larger than the values obtained before. I proceeded to check for duplicates, identifying 65 repeated values among the 1000 computed t-statistics. The presence of duplicates reduce the effective number of unique permutations and could introduce slight variability in the estimated p-value.

2.2 Single threshold test for multiple comparisons correction

2.2.a

In this section, I computed the two-sample t-statistic for each voxel using the GLM model $Y = X_1\beta_1 + X_2\beta_2 + e$. In analyzed only the voxels with a non-zero value in the mask and I found that the maximum observed t-statistic across all voxels was max_val $= 6.5294$.

2.2.b

Here, I generated all possible permutations of group labels while preserving sample sizes and for each permutation I recorded the maximum t-statistic, creating the empirical distribution maximum t-statistic shown in *Figure 3*.

2.2.c

I computed the multiple-comparison-corrected p-value by finding the proportion of permutations with a maximum t-statistic greater than the original value, obtaining $p_{\text{corrected}} = 9.1841 \times 10^{-2}$. This result, that is larger than typical significance thresholds, indicates that the we cannot reject H_0 at the 95% confidence level.

2.2.d

I determined the t-statistic threshold corresponding to $p = 0.05$ computing the $95th$ percentile of the empirical distribution and obtaining p_value_threshold $= 6.9383$, which is higher than the value found in $2.2.a$. This is expected, considering that the p-value obtained in $2.2.b$ is greater than 0.05.

3 Tables and Figures

$$X = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 \end{bmatrix}^T$$

Table 1: Design Matrix $[40x2]$ of $Y = X_1\beta_1 + X_2\beta_2 + e$.

1. Problem Definition	2. Orthogonality Condition of Residuals	3. Determination of β
$\hat{Y} = P_X Y$	$e = Y - \hat{Y}$	$\beta = (X^T X)^{-1} X^T Y$
$\hat{Y} = X\beta$	$X^T e = 0$	$\hat{Y} = X(X^T X)^{-1} X^T Y$
$P_X Y = X\beta$	$X^T(Y - X\beta) = 0$	$P_X = X(X^T X)^{-1} X^T$
	$X^T Y = X^T X\beta$	

Table 2: Derivation of P_X

1. Idempotency: $P_X^2 = P_X$	**2. Symmetry:** $P_X^T = P_X$
$P_X^2 = (X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T)$	$P_X^T = (X(X^T X)^{-1} X^T)^T$
$P_X^2 = X(X^T X)^{-1}(X^T X)(X^T X)^{-1} X^T$	$P_X^T = (X^T)^T (X^T X)^{-T} X^T$
Since $(X^T X)^{-1}(X^T X) = I_p$:	Since $(X^T)^T = X$ and $(X^T X)^{-T} = (X^T X)^{-1}$:
$P_X^2 = X I_p (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = P_X$	$P_X^T = X(X^T X)^{-1} X^T = P_X$

Table 3: Demonstration of P_X key properties

1. Idempotency: $R_X^2 = R_X$	**2. Symmetry:** $R_X^T = R_X$
$R_X^2 = (I - P_X)(I - P_X)$	$R_X^T = (I - P_X)^T$
$R_X^2 = I - 2P_X + P_X^2$	$R_X^T = I^T - P_X^T$
$R_X^2 = I - 2P_X + P_X$	$R_X^T = I - P_X$
$R_X^2 = I - P_X = R_X$	$R_X^T = R_X$

Table 4: Demonstration of R_X key properties

$$e = Y - X\beta$$

Minimize the sum of squared residuals

$$S(\beta) = e^T e = (Y - X\beta)^T (Y - X\beta)$$

$$S(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X\beta$$

$$\frac{\partial S(\beta)}{\partial \beta} = -2X^T Y + 2X^T X\beta$$

$$-2X^T Y + 2X^T X\beta = 0$$

$$X^T X\beta = X^T Y$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Table 5: Derivation of the general formula for estimating the model parameters of any GLM.

$$Y = X\beta + e$$

$$P_X Y = P_X(X\beta + e)$$

$$P_X Y = P_X X\beta + P_X e$$

$$P_X Y = X\beta + P_X e$$

$$P_X Y = \hat{Y} = X\hat{\beta}$$

$$X\hat{\beta} = X\beta + P_X e$$

$$P_X e = X\hat{\beta} - X\beta$$

$$P_X e = X(\hat{\beta} - \beta)$$

Table 6: Relation between $P_x e$ and $\hat{\beta} - \beta$.

$$e = Y - X\beta$$

$$R_X = I - P_X$$

$$e_\perp = R_X e = (I - P_X)e$$

$$e_\perp = (I - P_X)(Y - X\beta)$$

$$e_\perp = (I - P_X)Y - (I - P_X)X\beta$$

$$e_\perp = (I - P_X)Y - X\beta + X\beta$$

$$e_\perp = (I - P_X)Y$$

$$\hat{e} = (I - P_X)Y$$

$$e_\perp = \hat{e}$$

Table 7: Projection of e onto $C(X)^\perp$ and its relationship with $\hat{e}$.

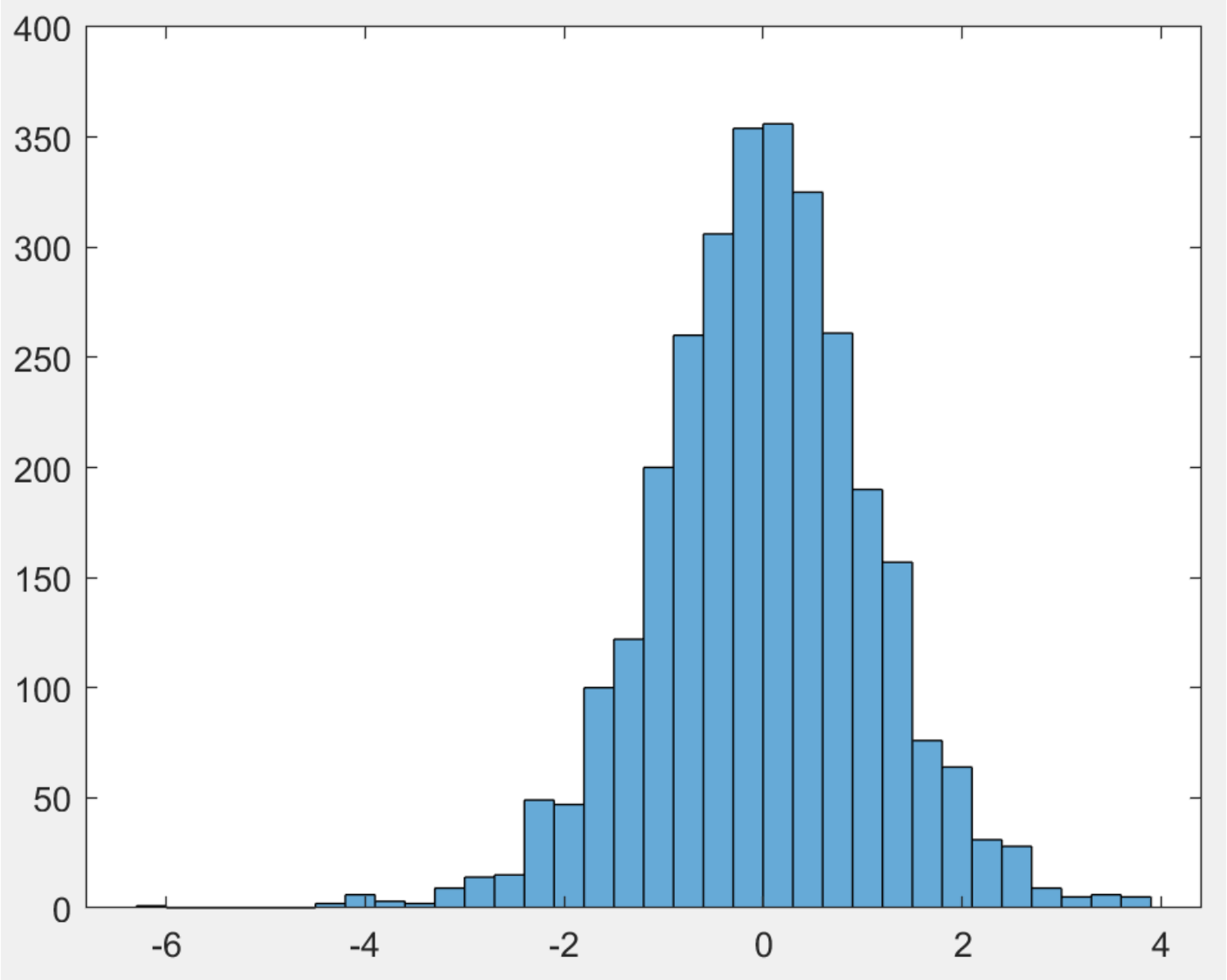

Figure 1: Empirical distribution of the t-statistic computed for all possible t-statistics from two sample groups of sizes $n_1 = 6$ and $n_2 = 8$

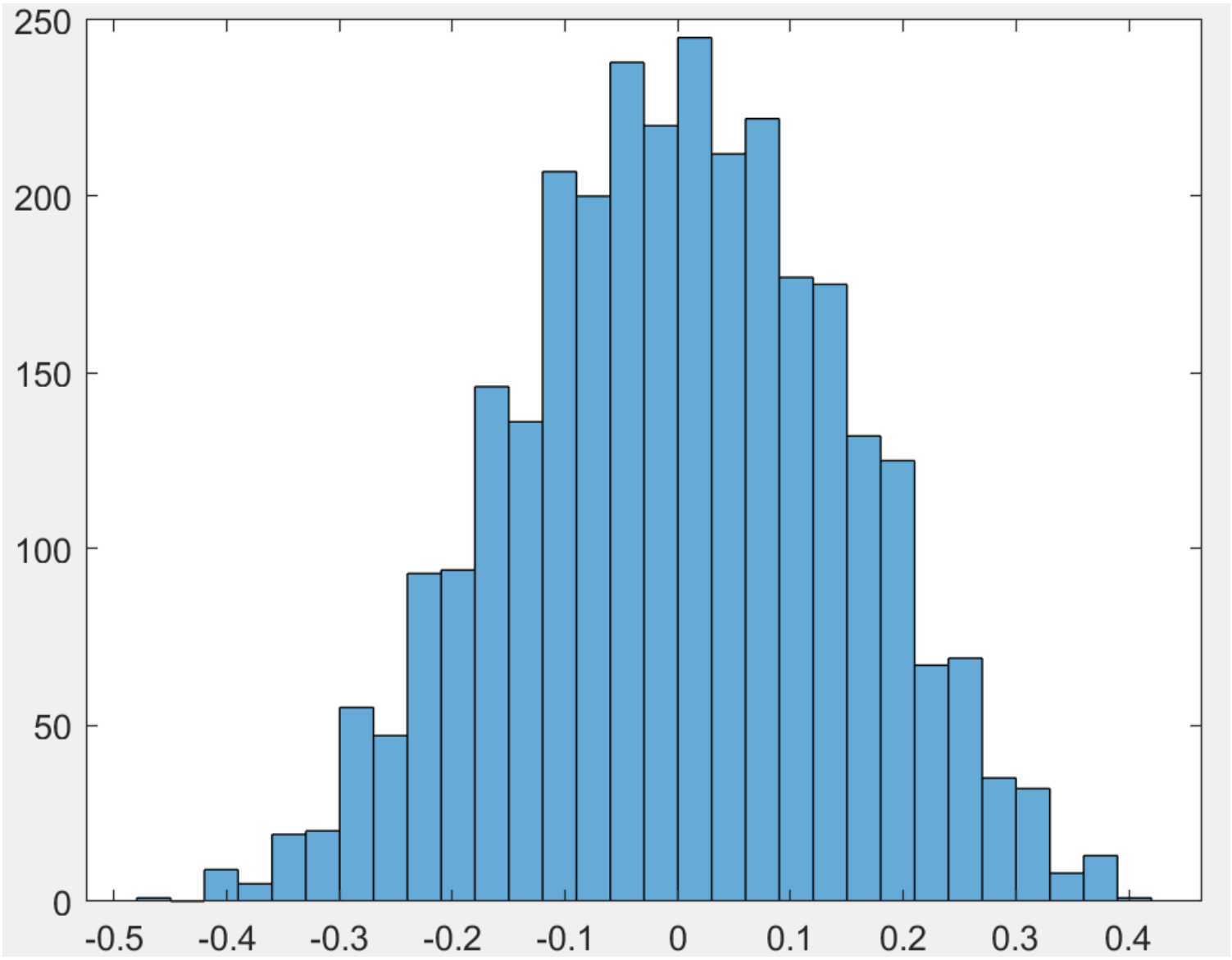

Figure 2: Empirical distribution of the mean difference computed across all possible mean differences for two sample groups of sizes $n_1 = 6$ and $n_2 = 8$

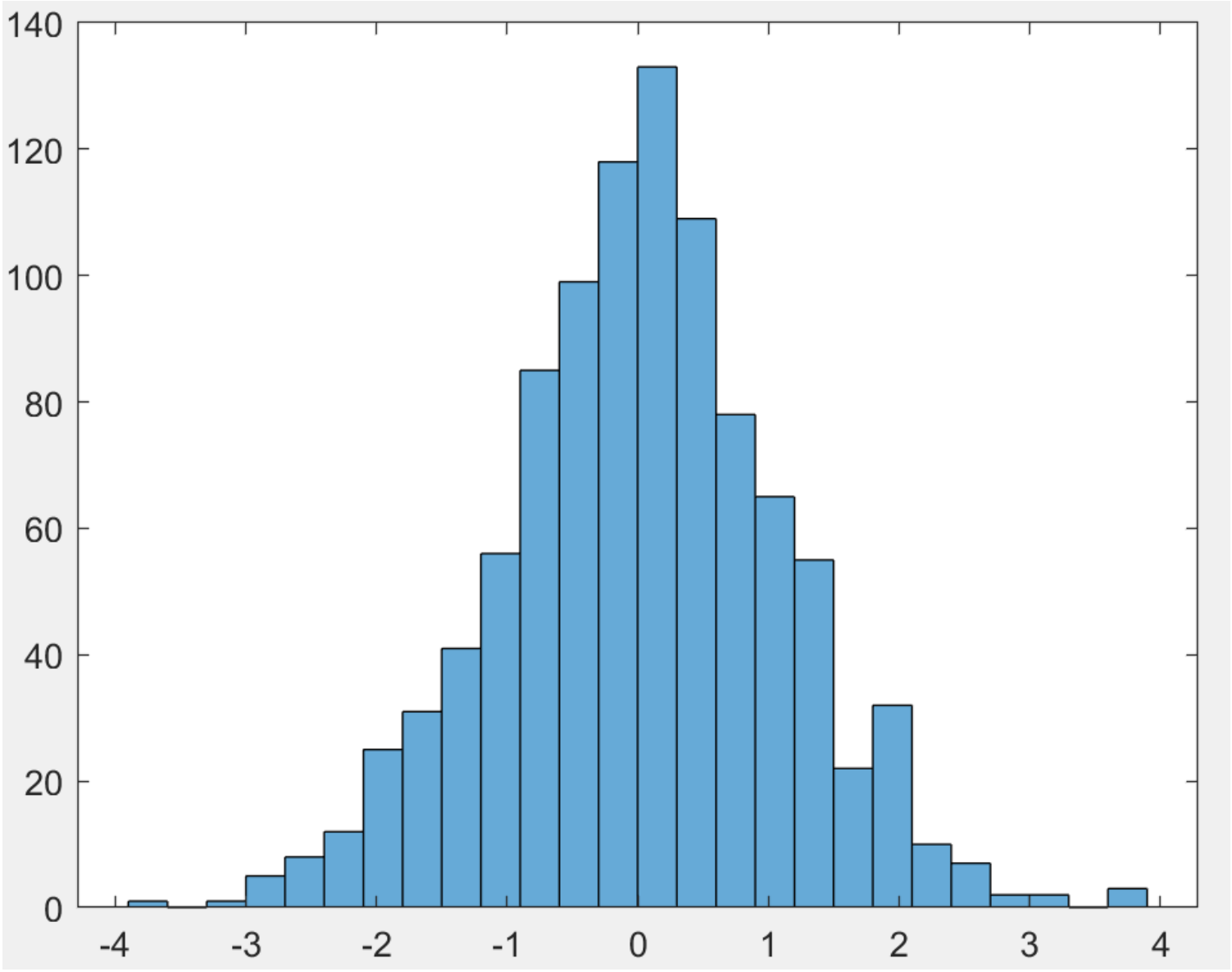

Figure 3: Empirical distribution of the t-statistic obtained from 1000 random permutations of two sample groups with sizes $n_1 = 6$ and $n_2 = 8$

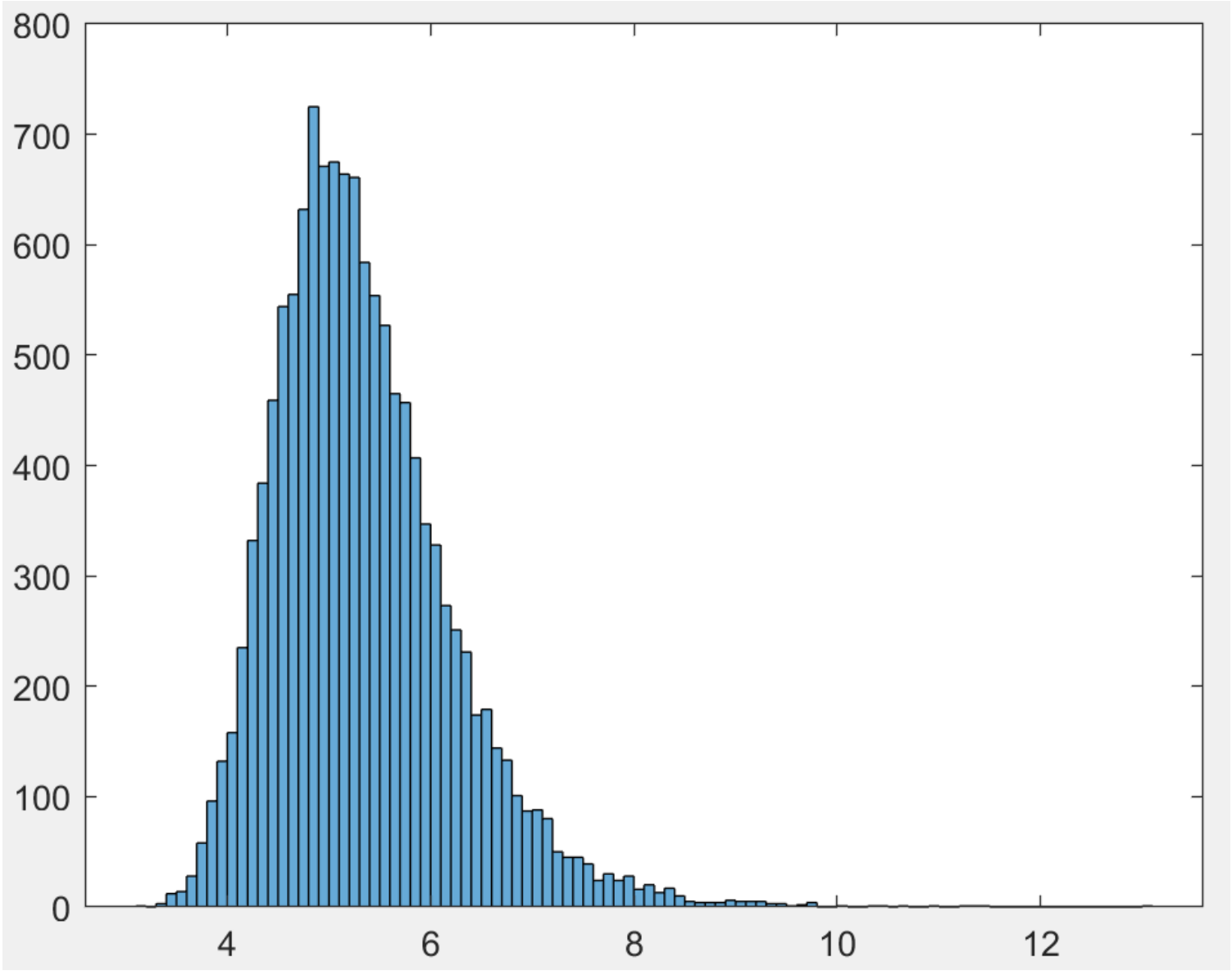

Figure 4: Empirical distribution of the maximum t-statistic derived from the fractional anisotropy (FA) maps of two groups, each consisting of 8 subjects