CM124/224, Fall 2023
Problem Set 1: Statistics and Multiple Testing
Due Oct 27, 2023 at 11:59pm PST

YAMAN YUCEL 605704529

10/27/23

1 Testing Mendel's first law [12 pts]

We set up an experiment to test Mendel's first law *i.e.*, the two copies of an individual's genome are equally likely to be transmitted to the offspring. We are examining a SNP at which the individual carries different alleles on their maternal and paternal genome. Denote the two alleles at this SNP as 0 and 1.

We observe the allele carried by n offspring. The state of the allele in an offspring i is given by a Bernoulli random variable $X_i \overset{iid}{\sim} \text{Ber}\,(p)\,, i \in \{1, \ldots, n\}$. Here p is the probability that a gamete inherits a 1 allele. Mendel's first law hypothesizes that $p = \frac{1}{2}$.

(a) Write the likelihood of p (Hint: to write the likelihood, we need the probability of a Bernoulli random variable X: $P(x) = p^x(1-p)^{(1-x)}$).

Solution: $\text{L(p)} = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$

$$= p^{\sum_{i=1}^{n}(x_i)}(1-p)^{\sum_{i=1}^{n}(1-x_i)}$$

(b) Write the log-likelihood of p.

Solution: $\text{l(p)} = \log(\text{L(p)}) = \sum_{i=1}^{n} \log\left(p^{x_i}(1-p)^{1-x_i}\right)$

$$= \sum_{i=1}^{n}(\log{(p)}x_i + \log{(1-p)}(1-x_i))$$

$$= \log{(p)}\sum_{i=1}^{n}(x_i) + \log{(1-p)}\sum_{i=1}^{n}(1-x_i)$$

(c) Show that the maximum likelihood estimator of p, $\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \overline{x}$.

Solution: $\frac{\partial l(p)}{\partial p} = \frac{\sum_{i=1}^{n}(x_i)}{p} - \frac{\sum_{i=1}^{n}(1-x_i)}{1-p} = 0$

$$=> \frac{\sum_{i=1}^{n}(x_i)}{p} = \frac{\sum_{i=1}^{n}(1-x_i)}{1-p}$$

$$=> \frac{1-p}{p} = \frac{n - \sum_{i=1}^{n}(x_i)}{\sum_{i=1}^{n}(x_i)}$$

$$=> \frac{1}{p} - 1 = \frac{n - \sum_{i=1}^{n}(x_i)}{\sum_{i=1}^{n}(x_i)}$$

$$=> \frac{1}{p} = \frac{n - \sum_{i=1}^{n}(x_i) + \sum_{i=1}^{n}(x_i)}{\sum_{i=1}^{n}(x_i)}$$

$$=> \hat{p} = \frac{\sum_{i=1}^{n}(x_i)}{n}$$

(d) Write the likelihood-ratio test statistic for testing $H_0 : p = \frac{1}{2}$ vs $H_1 : p \neq \frac{1}{2}$.

Solution:

$$\begin{aligned}
\lambda(\mathbf{X}) &= \frac{L(\frac{1}{2})}{L(\hat{p}_{MLE})} \\
&= \frac{(\frac{1}{2})^n}{(\hat{p}_{MLE})^{\sum_{i=1}^n (x_i)}(1-\hat{p}_{MLE})^{\sum_{i=1}^n (1-x_i)}} \\
&= \frac{(\frac{1}{2})^n}{(\frac{\sum_{i=1}^n (x_i)}{n})^{\sum_{i=1}^n (x_i)}(1-\frac{\sum_{i=1}^n (x_i)}{n})^{\sum_{i=1}^n (1-x_i)}}
\end{aligned}$$

(e) If we observe all alleles of type 1 across 5 gametes what is the exact p-value of the LRT statistic? Using this p-value, would you reject the null hypothesis at a significance level of 0.05?

Solution: First, I have computed the probability mass function of $\lambda(\mathbf{X})$, also LRT does not change due to the order of observation. Therefore, $\lambda(X = (1,0,0,0,0)) = \lambda(X = (0,1,0,0,0))$

λ(x = type 1 allele is not observed = (0,0,0,0,0)):

$$\hat{p}_{MLE} = 0 => \lambda(x) = \frac{1}{32} = 0.03125$$

This event occurs $\binom{5}{0} = 1$ times
λ(x = type 1 allele is observed once = 4 zeros 1 one):

$$\hat{p}_{MLE} = \frac{1}{5} => \lambda(x) = 0.3815 > 0.03125$$

This event occurs $\binom{5}{1} = 5$ times
λ(x = type 1 allele is observed twice = 3 zeros 2 ones:

$$\hat{p}_{MLE} = \frac{1}{5} => \lambda(x) = 0.9042 > 0.03125$$

This event occurs $\binom{5}{2} = 10$ times
λ(x = type 1 allele is observed third times = 2 zeros 3 ones:

$$\hat{p}_{MLE} = \frac{1}{5} => \lambda(x) = 0.9042 > 0.03125$$

This event occurs $\binom{5}{3} = 10$ times
λ(x = type 1 allele is observed four times = 1 zero 4 ones :

$$\hat{p}_{MLE} = \frac{1}{5} => \lambda(x) = 0.3815 > 0.03125$$

This event occurs $\binom{5}{4} = 5$ times
λ(x = type 1 allele is observed five times = (1,1,1,1,1):

$$\hat{p}_{MLE} = \frac{1}{5} => \lambda(x) = \frac{1}{32} = 0.03125$$

This event occurs $\binom{5}{5} = 1$ time
Then $P_{p=\frac{1}{2}}(\lambda(\mathbf{X}) \leq \frac{1}{32}) = \frac{1}{16} = \mathbf{0.0625}$
p-value of the LRT statistic is 0.0625, thus we do not reject the null hypothesis at a significance level $\alpha = 0.05 \leq 0.0625 = p_{value}$.

(f) Use the asymptotic distribution of the likelihood-ratio test statistic to compute the asymptotic p-value of the LRT statistic when we observe all allele type 1 for $n = 5$. Using this p-value, would you reject the null hypothesis at a significance level of 0.05?

Solution: $p(x) = P_{p=\frac{1}{2}}(\lambda(\mathbf{X}) \leq \frac{1}{32})$

$$P_{p=\frac{1}{2}}(-2\log(\lambda(\mathbf{X})) \geq -2\log(2^{-5}))$$

$$P_{p=\frac{1}{2}}(-2\log(\lambda(\mathbf{X})) \geq 10\log(2)) = 0.00847$$

Asymptotic p-value is 0.00847, thus we reject the null hypothesis at a significance level $\alpha = 0.05 \geq 0.00847 = p_{value}$

2 Multiple hypothesis testing [6 pts]

A study examines $m = 3226$ genes across two conditions. For each gene, the study aims to test the null hypothesis that the expression level of the gene is the same across the two conditions. Applying a hypothesis test, the study finds 51 genes to differ in their expression level across the conditions. Of these 51 genes, 9 are known to be truly null. Among genes found to not differ, 2000 are known to be truly null.

(a) What are the false positives, false negatives, true positives and true negatives for these tests?

Solution:

		Decision	
		H_0	H_1
Truth	H_0	$TN = 2000$	$FP = 9$
	H_1	$FN = 1175$	$TP = 42$

True Negatives = 2000
True Positives = 42
False Negatives = 1175
False Positives = 9

(b) The sensitivity or power is defined as the fraction of true non-null hypotheses that are predicted to be non-null. The specificity is the fraction of null hypotheses that are predicted to be null.What are the sensitivity and specificity?

Solution:

$$\begin{aligned} \text{Sensitivity} &= \frac{TP}{FN + TP} = \frac{42}{1217} = 0.034 \\ \text{Specificity} &= \frac{TN}{FP + TN} = \frac{2000}{2009} = 0.995 \end{aligned}$$

3 Data analysis [15 pts]

In this problem, you will test the association of SNPs to a phenotype using permutations as well as asymptotic approximations and compare the two approaches.

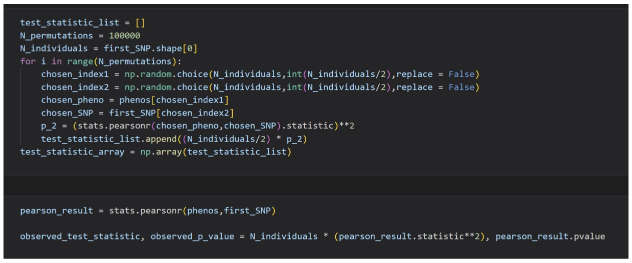
Provided is a data set of simulated phenotype (Y) for 250 individuals and a corresponding matrix of genotypes at 10 SNPs (G). We are interested in testing whether the genotype is associated with the phenotype in this data. To determine this, we will use the following test statistic.

$$T_i = N\rho^2(Y, G_i)$$

Here $\rho^2(Y, G_i)$ refers to the squared Pearson correlation coefficient between phenotype and genotype at the i'th SNP and N is the number of individuals.

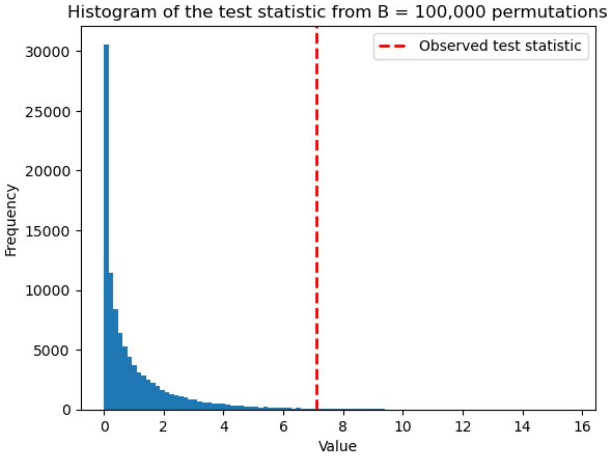
(a) Design and implement a permutation test for the first SNP (column 1 of the genotype matrix G). Plot the observed test statistic T_1 as well as a histogram of the test statistic from $B = 100,000$ permutations.

Solution:
Following script is written to perform permutation test for the first SNP. Firstly, I have randomly selected 125 (N/2) elements from phenos and first SNP, then computed the given statistic. This procedure is performed 100,000 times and the results are plotted in the following histogram.
$T_1 = 7.123297$



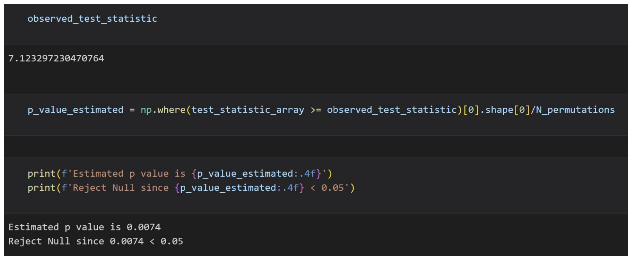
```
test_statistic_list = []
N_permutations = 100000
N_individuals = first_SNP.shape[0]
for i in range(N_permutations):
    chosen_index1 = np.random.choice(N_individuals,int(N_individuals/2),replace = False)
    chosen_index2 = np.random.choice(N_individuals,int(N_individuals/2),replace = False)
    chosen_pheno = phenos[chosen_index1]
    chosen_SNP = first_SNP[chosen_index2]
    p_2 = (stats.pearsonr(chosen_pheno,chosen_SNP).statistic)**2
    test_statistic_list.append((N_individuals/2) * p_2)
test_statistic_array = np.array(test_statistic_list)
```

```
pearson_result = stats.pearsonr(phenos,first_SNP)

observed_test_statistic, observed_p_value = N_individuals * (pearson_result.statistic**2), pearson_result.pvalue
```



(b) What is the p-value of T_1 estimated by permutations?

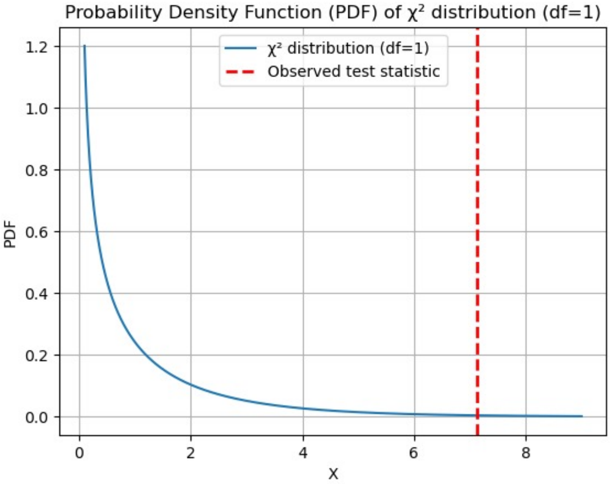
Solution: Estimated p-value of $T_1 = 0.0074 \leq 0.05$, therefore test rejects the null hypothesis.



```
    observed_test_statistic

7.123297230470764

    p_value_estimated = np.where(test_statistic_array >= observed_test_statistic)[0].shape[0]/N_permutations

    print(f'Estimated p value is {p_value_estimated:.4f}')
    print(f'Reject Null since {p_value_estimated:.4f} < 0.05')

Estimated p value is 0.0074
Reject Null since 0.0074 < 0.05
```

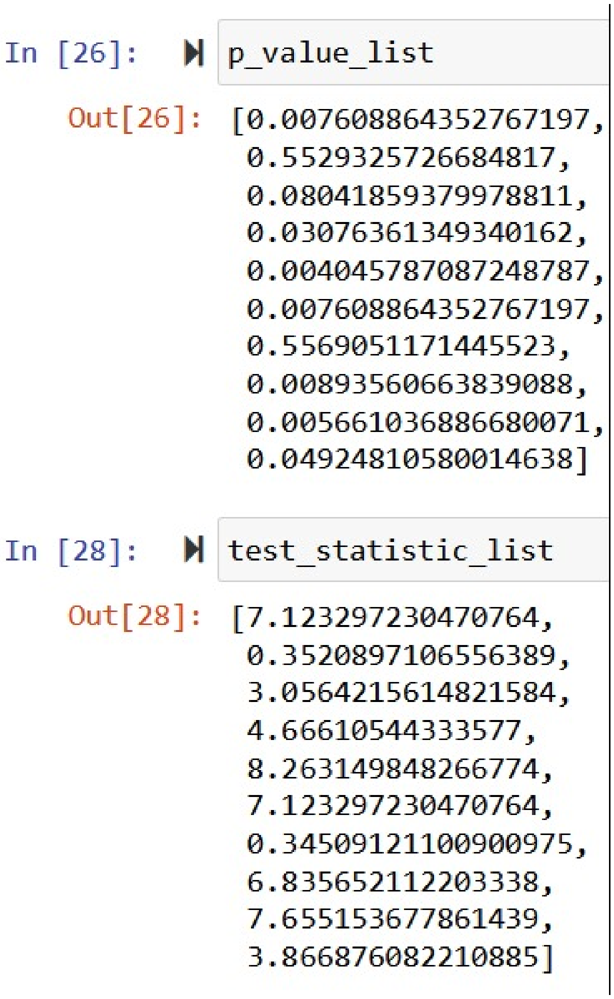
(c) The test statistic T_1 asymptotically follows a χ^2 with one degree of freedom under the null. Plot the probability density function for a chi squared with one degree of freedom. What is the p-value of T_1 based on the chi-square approximation?

Solution: $p - value = P(\chi^2 \ \geq 7.12) = 0.0076 \leq 0.05$, thus test reject the null hypothesis.



(d) We would now like to test each of the 10 SNPs for association with the phenotype but want to control the FWER at level 0.05. To do this, we first compute test statistics $(T_1, T_2, \ldots, T_{10})$ for each of the 10 SNPs and then compute p-values $(p_1, p_2, \ldots, p_{10})$ assuming the chi-squared distribution as in part (c). We propose to reject the null hypothesis that SNP i is not associated with phenotype if $p_i < t$. Our goal is to pick t so that the FWER is < 0.05. One approach to control FWER is the Bonferroni procedure. What is the p-value threshold t at which we should reject each of the 10 p-values using the Bonferroni procedure?

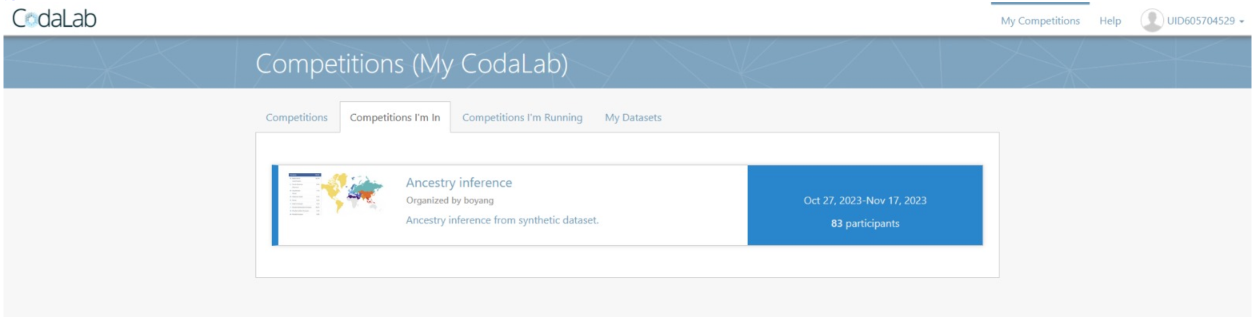
Solution: $t = \frac{0.05}{10} = 0.005$
Using the Bonferroni procedure, test only rejects 5th SNP since $0.00404 < 0.005$. Associated p values and test statistics are available below.



```
In [26]:   p_value_list

Out[26]: [0.007608864352767197,
          0.5529325726684817,
          0.08041859379978811,
          0.03076361349340162,
          0.004045787087248787,
          0.007608864352767197,
          0.5569051171445523,
          0.00893560663839088,
          0.005661036886680071,
          0.04924810580014638]

In [28]:   test_statistic_list

Out[28]: [7.123297230470764,
          0.3520897106556389,
          3.0564215614821584,
          4.66610544333577,
          8.263149848266774,
          7.123297230470764,
          0.34509121100900975,
          6.835652112203338,
          7.655153677861439,
          3.866876082210885]
```

4 Setup your account for the course projects [2 pts]

Solution:



Competitions (My CodaLab)

Ancestry inference

Organized by boyang

Ancestry inference from synthetic dataset.

Oct 27, 2023-Nov 17, 2023

83 participants