Project 1

1.

A. We would expect 100 individuals in the case group, since the probability of disease is P(d) = .10. Then, N\*p = 1000 \* .10 = 100. We run a binomial distribution with a .1 probability of each trial being a success. We run the test 1000 times (one per individual) with one trial each person (one chromosome per person).

Pseudocode:

-We use the random binomial function and then sort the tests into 0s and 1s (allele vs no allele).

-To find the number of individuals in the case group, we count the number of 1s, which represent individuals with the disease.

B. We need to create 1000 individuals with 200 SNPs each. Each SNP has a .5 chance of being present (equal chance of being present or not present), except the first SNP. The first SNP should have a .95 chance of being present in 100 individuals, and a .25 chance of being present in the other 900 individuals.

Pseudocode:

-Create an array of 1000 rows and 200 columns, and fill each index with either 0 or 1 randomly (.5 chance of each).

-Then, we create two new arrays of lengths 100 and 900, where the 100 length array has each index being 1 with a .95 chance, and the 900 length array with a .25 chance.

-Replace the first column of every row with this array, which represents the causal SNP.

C. In order to compare this data against the null hypothesis with α = .05, we need to find the highest ncp among all of the SNPs. We can calculate ncp with the equation:

We will compare this statistic to α /M, where M is the number of SNPs we take data for. In this case, we have αs = .05 / 200. Then, in order to reject the null hypothesis, the absolute value of the highest ncp must be greater than or equal to the absolute value of φ-1(αs).

Pseudocode:

Loop over all SNPs:

-Run the calculation for the ncp over all SNPs. For each SNP, check the statistic against the current maximum ncp found.

Note: The first SNP will nearly always be the most significant SNP

-If it is larger than the current maximum, overwrite the current maximum with the new maximum ncp.

-Next, we find the φ-1(αs). If our value is less than φ-1(αs) or greater than -φ-1(αs), the data is significant.

D. First, we must find the markers with correlations above .10. This can be done by comparing every SNP to every other SNP and saving correlations greater than .10. Then, we find the SNP with the most connections, add it to the tag group, and remove it and all SNPs correlated to it. This will be repeated until all SNPs are covered by the tag group.

Pseudocode:

-Find the correlations above .10 by running a correlation coefficient function on the SNPs and storing coefficients that are larger than .1 to a row and column array, where row[i] is correlated to col[i].

-Remove all duplicates from the arrays (i.e. 5 and 10 is equivalent to 10 and 5).

Loop until tag set covers all SNPs:

-Add the SNP with the most correlations (count occurrences in the arrays) to the tag set.

-Remove all correlations that contain this SNP. Save each correlated SNP removed in an array.

-To remove all SNPs it is correlated to, we must also remove all correlations that include any removed SNP (we saved this in the array in the previous step)

E.

To find the power, we must first find the ncp. This will be very similar to 1C. We can calculate ncp with the equation: . However, in this case, will always be .95-.25 instead of calculated from the allele frequency in the SNP. Additionally, α=.05 since we are only testing one SNP. Then, using the ncp, calculate power with: . To do the simulation, we create a large number of data sets and test for significance. We find the power by dividing the number of significant groups by the number of simulations. We expect both the analytic and simulated power to be near 1.0 because of the correlation in the first SNP to the case and control groups.

Analytic Pseudocode:

-Calculate ncp with equation above for the ncp statistic.

-Calculate power with the calculated ncp and the equation above for Power.

Simulation Pseudocode:

Loop a large number of times (we do 1000):

-Create a new data set

-Test for significance, if success, increment number of successes by 1

-divide number of successes by 1000

2.

A.

In order to have 20% of individuals correlated with τ = 1, we just need to have 20% of case individuals exactly identical and 20% of control individuals exactly identical. This can be done fairly easily by copying one individual 900 \* .2 – 1 = 179 times in the control group. Likewise, copy an individual 100 \* .2 – 1 = 19 times in the case group.

Pseudocode:

-Run code from 1B to create uncorrelated sample groups

-Take the first individual in the control group and copy over the next 179 individuals.

-Take the first individual in the case group and copy over the next 19 individuals.

-Run functions created in 1C, 1D, and 1E using the new case and control groups.

Differences: Both 2A and 1C will nearly always be significant because of how correlated the allele frequency in the first SNP is to the control and case groups. The correlated data will have a much smaller tag set since there will be a much larger amount of correlations above .10, since 20% of individuals are exactly alike. The analytic power will be the same value, since power calculation does not take into account the case and control group data. The simulated powers should be approximately 1, because of the large correlation between the minor allele frequency and to the control and case groups.

B.

In order to have 50% of individuals correlated with τ = 1, we just need to have 50% of case individuals exactly identical and 50% of control individuals exactly identical. This can be done fairly easily by copying one individual 900 \* .5 – 1 = 449 times in the control group. Likewise, copy an individual 100 \* .5 – 1 = 49 times in the case group.

Pseudocode:

-Run code from 1B to create uncorrelated sample groups

-Take the first individual in the control group and copy over the next 449 individuals.

-Take the first individual in the case group and copy over the next 49 individuals.

-Run functions created in 1C, 1D, and 1E using the new case and control groups.

Differences: Same as 2A except for the tag set. The correlated data will have a much smaller tag set (generally smaller than 2A) since there will be a much larger amount of correlations above .10, since 20% of individuals are exactly alike.

C.

In order to have 20% of individuals correlated with τ = .5, we take advantage of the fact that parents and their children, as well as siblings from the same parents have a .5 correlation to one another. This is tue because a parent and child will generally share 50% of their DNA (50% from father and 50% from mother). Using this, we can create correlated individuals by “breeding” two individuals. We choose two individuals from the control group, and breed them 900 \* .2 – 2 = 178 times. These 180 individuals will have approximately a .5 correlation to one another. Likewise, choose two individuals from the control group, and breed them 100 \* .2 – 2 = 18 times.

Pseudocode:

-Run code from 1B to create uncorrelated sample groups

-Take the first two individual in the control group. Breed them 178 times and place into control group.

-Take the first two individual in the case group. Breed them 18 times and place into control group.

-Run functions created in 1C, 1D, and 1E using the new case and control groups.

Breeding Pseudocode:

Loop over each SNP:

-For each SNP, if both parents have it, the child will have it. If neither parent has it, the child will not.

-If only one parent has it, but the other does not, randomly decide if the child will or will not have it.

Differences: Same as above, except for tag set. The correlated data will have a smaller tag set than 1C, but will be larger than 2A and 2B since there is some correlation between 20% of the individuals, but the correlations are only 50% instead of 100% identical.