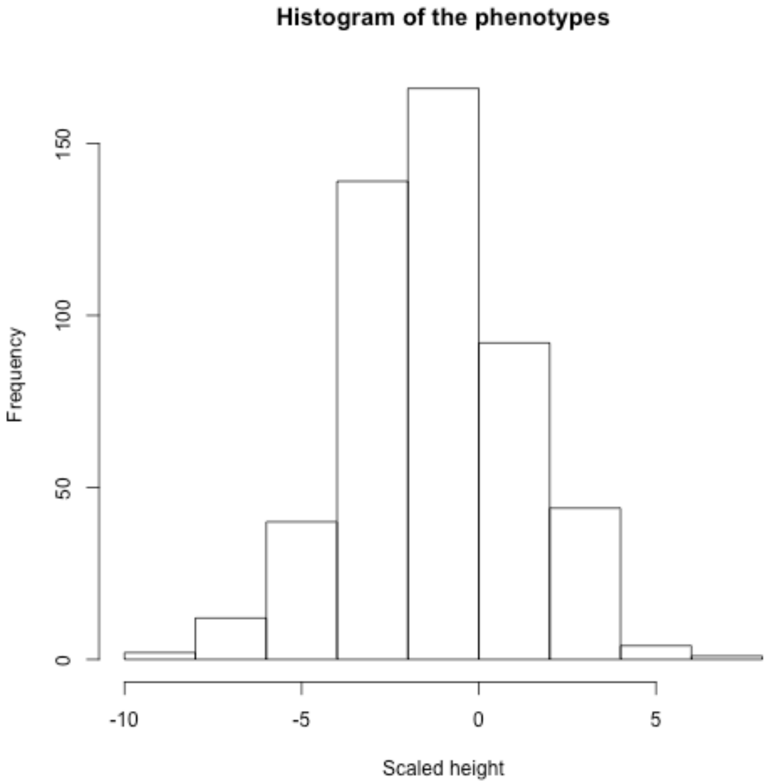
Midterm Exam

Dan Jin. Oct. 12, 2012

Problem 1:
The sample size **(n) = 500**; the number of genotypes per individual **(N) = 1194**.

Problem 2:

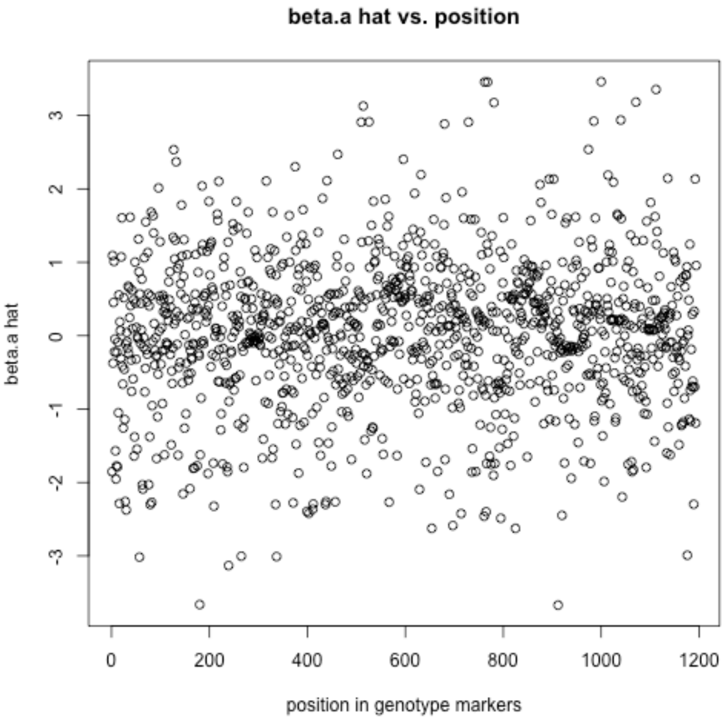
1. Histogram of the phenotypes



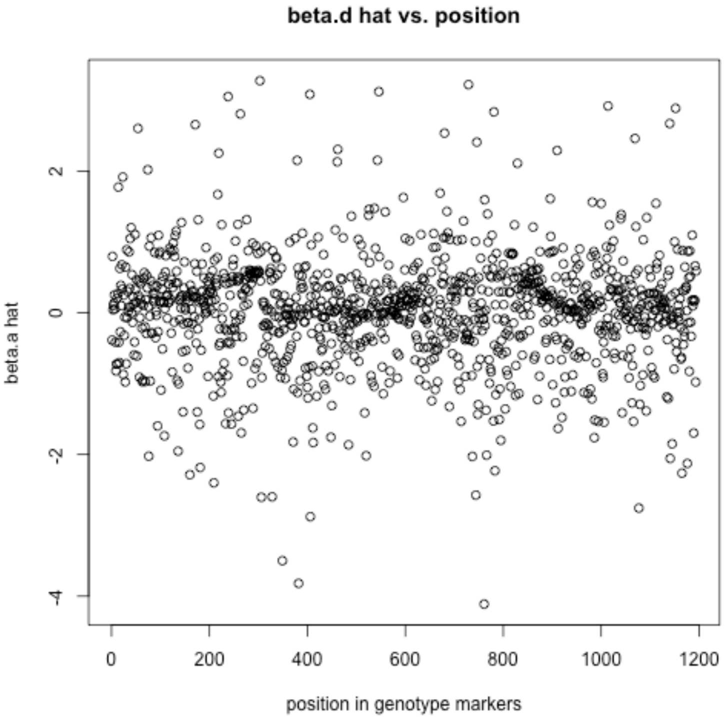
2. **Yes**, these phenotypes are approximately normally distributed.
3. Because the F-test we're using requires that the data be normally distributed.

Problem 3:

1. Write R code to calculate the MLE($\hat{\beta}$): Please refer to "MidtermScript.r".
2. Plot the value of $\hat{\beta}_a$ vs. genotype markers (Please refer to "MidtermScript.r" for R code):



3. Plot the value of $\hat{\beta}_d$ vs. genotype markers (Please refer to "MidtermScript.r" for R code)



4. For the majority of the genotypes, we expect the true **β_a = 0** and **β_d = 0**. **Because** the majority of the genotypes are neither causal genotype(s) nor tightly linked with the true causal genotype(s), and should not correlated with the phenotypes.

Estimator is a statistic on a sample, defined to return a value that represents our best evidence for being the true value of a parameter. **The goal of constructing an estimator** is to use the sample to determine the "true" parameter value that describes the outcomes of our experiment.

The reason why none of the $\widehat{\boldsymbol{\beta}}_{\boldsymbol{a}}$ = 0 and $\widehat{\boldsymbol{\beta}}_{\boldsymbol{d}}$ = 0 is that the sample size is limited. We are not able to observe the entire sample space. So the estimators (e.g. $\hat{\beta}_a$ and $\hat{\beta}_d$) are calculated based on limited sample, therefore are not equal to the true value.

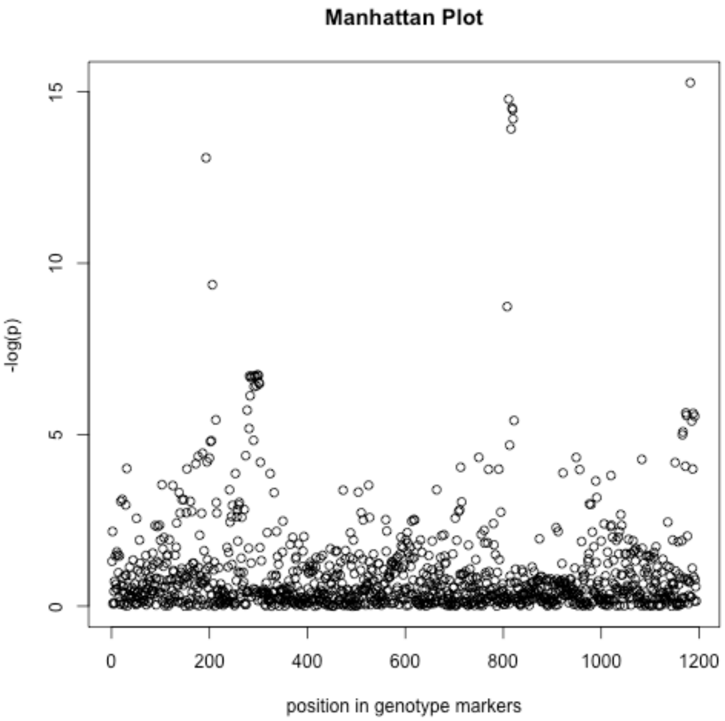
Problem 4:

1. Write R code to calculate the p values testing the H0: Please refer to "MidtermScript.r".
2. The alternative hypothesis is: **H_A : βa ≠ 0 ∪ βd ≠ 0**.
3. **p value** is the probability of obtaining a value of a statistic T(x), or more extreme, conditional on H0 being true.

   In this case, each p value represents the probability of obtaining a value (= F statistics), or more extreme, conditional on H_0 : $\beta_a = 0$ and $\beta_d = 0$.

Problem 5:

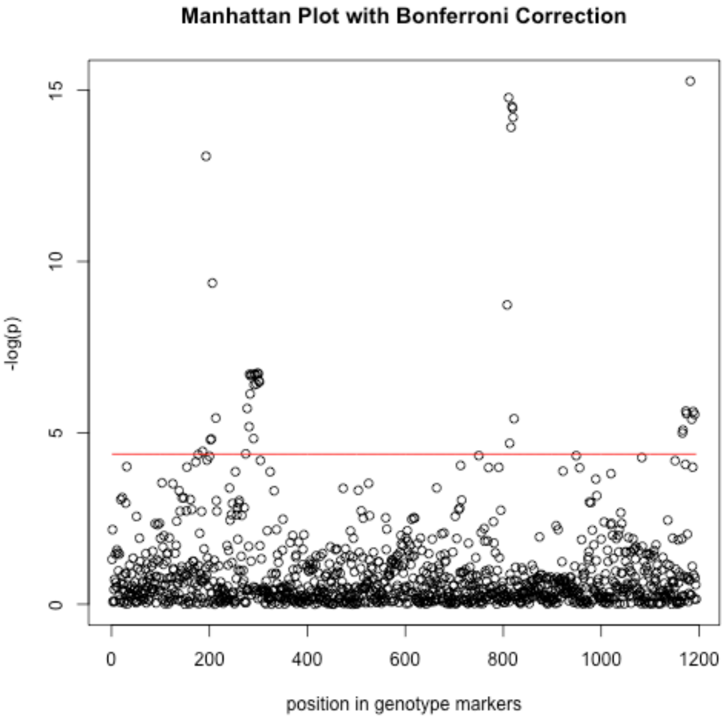
Manhattan plot (Please refer to "MidtermScript.r" for R code):



Problem 6:

1. Bonferroni corrected type I error: $\boldsymbol{\alpha_B} = \frac{\boldsymbol{\alpha}}{\boldsymbol{N}} = \frac{\mathbf{0.05}}{\mathbf{1194}} = \mathbf{4.187605e-05}$

2. Manhattan plot with Bonferroni threshold (Please refer to "MidtermScript.r" for R code):
Note: the threshold on Manhattan plot should be $–log_{10}$ ($α_B$) = 4.378034



3. **Definition of Type I error:** Type I error is the probability of incorrectly rejecting H_0 when it is true.
4. **The multiple testing problem:** In a GWAS, we perform N hypothesis tests, and we would expect to incorrectly reject the H_0 (N*Type I error) times. If N were large, we would make lots of error. So we want to use a Bonferroni corrected Type I error, which sets the Type I error for the entire GWAS and reduce the then number of error.
5. **Definition of Type II error:** Type II error is the probability of incorrectly accepting (or not rejecting) H_0 when it is false. **We cannot control Type II error** because it depends on the actual parameter value, which we don't know.
6. **The trade-off between Type I error and Type II error:** For a give sample, decreasing Type I error will increase Type II error, and vice versa.
7. Applying a Bonferroni correction decreases the Type I error of a test and therefore **increases** the Type II error of a test.
8. **Definition of power**: power equals to 1-Type II error. It represents the probability of correctly rejecting H_0 when it is false. We cannot control power because like Type II error, it also depends on the actual parameter value, which we don't know.

9. **The trade-off between Type I error and power:** For a given sample, decreasing Type I error will decrease power, and vice versa.
10. Applying a Bonferroni correction decreases the Type I error of a test and therefore **decreases** the power of a test.

Problem 7: (Please refer to "MidtermScript.r" for R code)

1. List of all genotype markers with their corresponding p values:

```
$hit.position
[1]  186  193  202  204  206  213  274  277  281  282  283  284
287  290
[15]  291  292  294  296  298  300  301  302  808  811  813  816
818  819
[29]  820  822 1166 1167 1173 1174 1182 1185 1188 1191

$hit.pval
 [1]  4.457882 13.070888  4.798361  4.823398  9.372000  5.431085
4.388118
 [8]  5.712318  5.178458  6.710362  6.134569  6.665158  6.701824
4.833773
[15]  6.401800  6.704053  6.704053  6.423896  6.742694  6.727083
6.521027
[22]  6.484797  8.734447 14.781594  4.694469 13.914018 14.524456
14.476910
[29] 14.205841  5.413049  4.995273  5.081697  5.635441  5.560703
15.262291
[36]  5.394543  5.619389  5.536150
```

2. Diveide this list into sets: I divided them into **4 distinct sets**.

```
[[1]]
[1] 186 193 202 204 206 213

[[2]]
 [1] 274 277 281 282 283 284 287 290 291 292 294 296 298 300 301
302

[[3]]
[1] 808 811 813 816 818 819 820 822

[[4]]
[1] 1166 1167 1173 1174 1182 1185 1188 1191
```

3. List the most significant marker in each sets with the associated p values:

```
$position
[1]  193  298  811 1182

$pval
[1] 8.493990e-14 1.808449e-07 1.653509e-15 5.466501e-16
```

4. I divided them into these sets **based on the distance between two adjacent hits:**
   Step 1: Measured the distances between two adjacent hits.
   Step 2: Set a threshold for clustering by calculating the average distance between two adjacent hits.
   Step 3: If the distance between two adjacent hits is smaller than the threshold, they will be grouped into the same set. Otherwise, they will be put into distinct sets.

   This method depends on the following assumptions:

   a) The marker numbers reflect their positions (their order) in the genome.
   b) The distance between two hits from two different sets should be much larger than the distance from two hits within a set.
   c) There should not be too many hits, so most distance I measured should be between two hits from the same set and should have a small value.

Problem 8:

1. Calculate the MAF for the most significant markers in each set (Please refer to “MidtermScript.r” for R code):

```
$MAF
[1] 0.222 0.288 0.430 0.259
```

2. The relationship between MAF and power: **Power tends to increase as the MAF increases.** Because *lower MAF* makes this locus “look” like *correlated with the phenotype*, although it is actually not. This will make us reject H_0 when we should not, and therefore *increase Type II error* and *decrease power*.

Problem 9: (Please refer to “MidtermScript.r” for R code)

1. The most significant genotype marker in GWAS, both marker number and the p value:

```
$position
[1] 1182

$pval
[1] 5.466501e-16
```

2. Correlation between its X_a and the X_a for either the marker that is in the column to the left or right:

```
$position.left
[1] 1112
$corr.left
[1] -0.02848561

$position.right
[1] 1185
```

```
$corr.right
[1] 0.568552
```

3. Correlation between its X_a and the X_a for marker number 409

```
$corr.409
[1] -0.0005280443
```

4. The correlation between the X_a of the most significant genotype marker in GWAS and the X_a of marker number 1185 is higher, because they are physically closer in the genome, therefore more tightly linked.

   If two markers are in **Linkage Disequilibrium**, it means 1) these two markers are physical linked on a chromosome (Linkage) AND 2) they are not in H-W equilibrium (Disequilibrium). **This matches our expectation for a GWAS.** We expect a tag in GWAS is linked with the true causal polymorphism. And due to this linkage, there should be fewer recombination events happened between them, causing disequilibrium.

Problem 10:

1. We generally consider a set of linked and contiguous markers with significant p values to indicate the position of a single causal polymorphism. But each significant marker in a contiguous set does not indicate a separate causal polymorphism, **because they are probably in LD with the same causal polymorphism**. On the other hand, due to LD, for a single causal polymorphism, we should be able to find a region (rather than a single point) where a set of markers around this causal polymorphism gives high correlation score with the phenotype (given enough markers in that region).
2. **A causal genotype** refers to a position in the genome where an experimental manipulation of the DNA produces an effect on the phenotype on average or under specified condition. GWAS is a mapping experiment. When we reject the H_0 we only assume we have located a position in the genome that contains a causal polymorphism. **The significant markers we found in GWAS (in most cases) are tags in LD with the causal genotypes.** They are not the loci that really cause the phenotype, but they are linked with the real causal genotype during recombination and **appear** to be correlated with the phenotype. **If we mutate these markers, the mutations will not affect the phenotype on average**.
3. Because these significant markers only indicate a certain polymorphism within this region has high correlation with the phenotype, but **correlation ≠ causality**. **If the phenotype we are interested in is correlated with a second phenotype, then the significant marker we found in the GWAS may actually indicate a causal polymorphism of that second phenotype.**

For example, in this case we are trying to find genetic loci that affect human height. But human height seems also correlated with human racial. European people have higher average height than Asian people. And European people also have lighter hair color. So, the significant markers we found in question 7 may actually indicate a causal polymorphism of hair color instead of human height.