Ryan Oliveira

Statistical Genomics (CROPS 545), Spring 2017 Homework #2

(1) **Hypothesis:** Stochastic imputation on missing values in the same marker data set will remain constant in accuracy as the missing rate increases. We hypothesize this based on the fact stochastic imputation uses a set random formula that does not use the non-missing data to determine its values.

Methods: Data set mdp_numeric.txt is used and has 281 individuals with 3093 SNPs. A stochastic imputation formula (see source code) is used and generated on each missing data set. The accuracy is calculated via a simple correlation coefficient between the imputed and original matrices. This process is repeated 30 times. The average, standard deviation, and number of repetitions are recorded.

Results: Missing data is successfully generated. The below table summarizes the actual count of missing values randomly generated vs. the goal. The actual rates are almost identical to the goal.

Set missing	5%	10%	25%	50%	75%
Actual missing	0.0504	0.100	0.251	0.500	0.750

Accuracy is listed below for each missing rate. Accuracy varies only by 0.001 between missing rates, lending support to our hypothesis its accuracy remains constant due to being a random value.

Set missing	**# Repetitions**	**Accuracy avg.**	**Accuracy SD**
5%	30	0.388	0.00418
10%	30	0.387	0.00298
25%	30	0.387	0.00184
50%	30	0.387	0.000983
75%	30	0.387	0.00128

A consistent downward trend is seen in the standard deviation with an increasing percentage set missing. This is attributed to the fact that averages of a higher sample size (e.g., more values to impute) tend to cluster more closely.

(2) **Hypothesis:** Similar to what is hypothesized in problem (1), as stochastic imputation randomly generates results without using informed data from the rest of the data set, its accuracy will not change regardless of the number of individuals sampled.

Methods: The missing rate is fixed to 25%. Stochastic imputation is performed on a randomly sampled proportion of the dataset set to either 25%, 50%, and 75% of the total individuals. This process is repeated 30 times each and reports the accuracy by correlation coefficient. The accuracy's average, standard deviation, and number of repetitions for each set of repetitions are recorded.

Results: Accuracy is listed as before. As hypothesized, the accuracy is invariant with differing sample sizes. Similar to the last sampling set, standard deviation decreases with larger sample sizes.

Set sampled	# Repetitions	Accuracy avg.	Accuracy SD
75%	30	0.386	0.00236
50%	30	0.387	0.00464
25%	30	0.387	0.00693

(3) **Hypothesis:** Using the K-Nearest-Neighbor imputation method on generated missing values, we hypothesize that as K is increased on the same market dataset, accuracy will increase due to the available of more training examples. Standard deviation is expected to increase alongside the average.

Methods: In each of twenty repetitions, approximately 25% of the reference data is randomly set to missing, and the K-Nearest-Neighbor (KNN) algorithm is used to impute the missing data. The accuracy is obtained by correlation coefficient for each run, and the average and standard deviation of the 20 repetitions are calculated. This process is performed for k set to 2, 5, 10, and 20.

Results: As hypothesized, the average accuracy (and its standard deviation) increases as more "nearest neighbors" are available for training.

K	# Repetitions	Accuracy avg.	Accuracy SD
2	20	0.538	0.000649
5	20	0.607	0.000988
10	20	0.626	0.00153
20	20	0.630	0.00148

(4) **Hypothesis:** When the methods are compared on the same marker dataset, we hypothesize that BEAGLE will have the highest imputation accuracy, followed by KNN, followed by stochastic imputation.

Methods: Three methods are used: stochastic imputation, K-Nearest-Neighbor, and BEAGLE. For each method, ten repetitions each are performed on a random 25% of missing data, and the accuracy is reported by correlation coefficient. At the end of the ten runs, the mean and standard deviation values from the accuracy vector is calculated.

Results: BEAGLE imputes the missing data with the highest accuracy and stochastic imputation the lowest. Standard deviation increases with the magnitude of the average accuracy.

Method	# Repetitions	Accuracy avg.	Accuracy SD
Stochastic	10	0.386	0.000562
KNN (k=10)	10	0.649	0.000713
BEAGLE	10	0.822	0.00125

(5) **Hypothesis:** Performing K-Nearest-Neighbor treating markers as "neighbors" and individuals as "attributes" will result in increased accuracy compared to the inverse in the tested data set, as the number of markers in this set vastly outnumbers the number of individuals.

Methods: Methods are repeated identically to #3, with the exception that the rows and columns of the data set are transposed, allowing the KNN test to treat markers as neighbors. The data from this transposed method is compared to the original data from #3, and a paired *t*-test is conducted.

Results: As expected, the greatly expanded number of markers allowed for greater accuracy when they were made available as "neighbors" for the KNN imputation method. In addition, the standard deviation is considerably less than those obtained when using individuals as neighbors. This decreased standard deviation is likely secondary to having a much larger number and therefore higher density of neighbors, allowing for more consistency with the KNN imputation method.

Method	Markers = Neighbors			Individuals = Neighbors		Compared
K	# Reps	Acc. avg.	Acc. SD	Acc. avg	Acc. SD	*p-value*
2	20	0.652	0.000163	0.538	0.000649	
5	20	0.680	<0.000001	0.607	0.000988	$p < 0.01$
10	20	0.692	<0.000001	0.626	0.00153	
20	20	0.690	<0.000001	0.630	0.00148	