B565: Homework 1, Due

1. In the “curse of dimensionality” R example done in class we let X be a d-dimensional vector of “runif” variables. This means the components of X are independent $\mathsf{Unif(0,1)}$ random variables. The R example computes the distance of X to the edge of the hypercube by

$$D = \frac{1}{2} - \max_{i=1\ldots d} |\frac{1}{2} - X_i|$$

   (a) The definition of independence says that for $X_1, \ldots, X_d$,

$$P(X_1 \in A_1, \ldots, X_d \in A_d) = \prod_{i=1}^{d} P(X_i \in A_i)$$

   What is the probability that a point is at least a away from an edge? That is what is $P(D > a)$ for $0 < a < \frac{1}{2}$.

   (b) Given any distance to edge δ, find d so that the probability of X lying within δ of the edge of the hypercube is at least .9. That is, find d so that $P(D < \delta) \geq .9$.

2. Consider the pairs plot of the “iris” data discussed in class. If we want to predict the iris species from the petal and sepal measurements, and had to eliminate one of the predictors, which would you eliminate based on the plot. Explain your choice clearly.

3. A clear jar contains some unknown number of beans. 100 randomly chosen people inspect the jar and make estimates, $n_1, \ldots, n_{100}$ of the true number of beans, n_{true}, with no knowledge of the others’ estimates. The sample average, $\bar{n} = \frac{1}{100}\sum_{i=1}^{100} n_i$, is computed.

   (a) Consider modeling $n_1, \ldots, n_{100}$ as a random sample from some distribution with mean μ and variance σ^2. Why or why not is this a reasonable model?

   (b) What is the approximate distribution of $\bar{n}$?

   (c) What is the approximate probability that $\bar{n}$ misses μ by more than 2 beans?

   (d) What can you say about the actual error, $\bar{n} - n_{\text{true}}$. Make clear any assumptions you are making.

4. Consider the trees data set which you can import by

```
> data(‘‘trees’’)
```

   This data set measures girth, height, and volume on a small collection of trees. Treat the data set as a matrix, using

```
> X = as.matrix(trees)
```

   The features (colummns) can be centered to have mean 0 using

```
> X = scale(X,center=TRUE,scale=FALSE)
```

   you may wish to use a pairs plot to see this has the desired effect.

   (a) **Using matrix calculations** compute the sample covariance matrix. Do not use an R package to do this.

   (b) Perform principal component analysis by computing the singular value decomposition (svd) of the sample covariance matrix.

   (c) Using the column of U associated with the smallest diagonal element, find an approximate linear model between the three centered variables.

5. Consider the “mystery.csv” data set available from Canvas. You can read this data set in with

```
> read.csv("mystery.csv")
```

if the data is in your working directory. Using principal components, find the effective dimension of the data — this is, dimension of the hyperplane that contains the data points. Using this result perform dimensionality reduction on the data matrix to give a smaller number of columns without any loss of information.

6. In your book, Chapter 2: problems 16,17, 25,26.