The biopsy dataset includes 569 observations collected by the University of Wisconsin in 1992 from fine needle aspiration biopsies on masses found in breast tissue. During this procedure, a hallow needle attached to a syringe is used to withdraw tissue or fluid from the mass. The variables in the dataset describe the characteristics of the cell nuclei obtained during the biopsies and the diagnosis. The variable are as follows:

- radius: the mean of distances from the center of the cell nuclei to the points on its perimeter
- texture: the mean texture of the nuclei, described by the spatial arrangement and variation of grey values observed
- perimeter: the mean distance around the nuclei
- area: the mean area of the nuclei
- smoothness: the mean of the local variation in radius lengths
- diagnosis: the diagnosis of the mass, where 0 = malignant, 1 = benign

I want to use the dataset to determine what variables are significant to the diagnosis of breast cancer. I also want to use a machine learning model to predict the diagnosis with high accuracy.

I suspect radius, perimeter, and area are highly correlated. Although a cell nucleus may not be perfectly round, I am considering the geometry of a circle. The perimeter, or circumference in this case, is $2\pi r$ and the area is πr^2. Thus, radius is causally correlated with perimeter and area. I used scatter plots to confirm the high correlation between variables. Because radius and perimeter can be used to describe area, I decided to omit them from any further analysis and only use area.

To decide what variables are significant, I used PMFs and CDFs of the remaining variables (texture, area, and smoothness) to compare the two diagnosis groups. The malignant diagnosis seems to be associated with larger values of these variables. To confirm there is a significant difference between the diagnosis groups, I used permutation hypothesis tests. The results prove statistical significance between malignant and benign diagnosis for all three variables.

The CDFs of texture, area, and smoothness have the shape of a lognormal distribution so the use of a lognormal model fits the data very well. Taking this into account, I used the log transformation of texture, area, and smoothness in a logistic regression model to predict diagnosis with 94% accuracy.

References

Fine Needle Aspiration (FNA) Biopsy of the Breast: Breast Aspiration. (n.d.). Retrieved August 06, 2020, from https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html

Singh Suwal, M. (2018, September 26) Breast Cancer Prediction Dataset. Retrieved July 07, 2020 from, www.kaggle.com/merishnasuwal/breast-cancer-prediction-dataset