# Data Description

## Overview of the Data Set

The data is a collection of the Breast Cancer Diagnostic Data Set collected by the University of Wisconsin-Madison. 569 observations were taken using digitized imaging of breast mass biopsy samples taken between 1989 and 1991. Biopsies were acquired using FNA, Fine Needle Aspiration

The data set was taken from UC Irvine’s Machine Learning repository.

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

Below is a description of the data set

DEPENDENT VARIABLE

Diagnosis: Binomial – B [benign] or M [malignant]

INDEPENDENT VARIABLES

Ten real-valued features are computed for each cell nucleus:

a) radius (distance from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

For each of these metrics the data set included

* Mean
* Standard Error
* Worst [for asymmetrical masses this is the “longest” or “largest” value]

## Initial Data Exploration

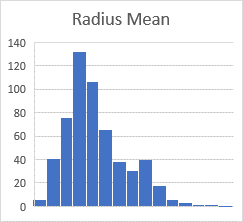
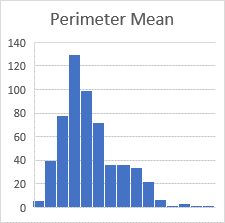
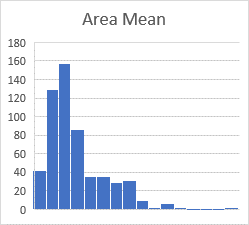
First

## Data Assumption Checks

**Sample Size**: The sample size is moderately sized with over 500 observations. Approximately 37.3% were Malignant and 62.7% were Benign.

**Independence**: Specific details as to patient selection were not available, so we will assume that independence were maintained. Points of possible concern would be a sample without proper diversity to represent the true population.

**Normality**: Normality assumptions were for the most part met. In cases where skew exists the CLT applied.

## Analysis & Correlation

**Radius, Perimeter, Area**

These are obviously highly correlated since they are all a function of the radius and Pi.

Upon initial visual examination there appears to be a correlation present between the measurements of the tumorous mass and the “malignant” vs “benign” diagnosis.

The line chart below shows the progression of measurements, from smallest to largest, for both radius and area, separated by diagnosis. Over all the malignant biopsies appear to be larger overall.

A box plot comparison also validates this initial observation

APPENDIX

**Code: Box Plot Analysis**

#Box Plot: Area Mean

boxplot(area\_mean ~ diagnosis,data=wdbc,

horizontal=TRUE,

names=c("Benign","Malignant"),

col=c("green","red"),

xlab="Area Mean", main="Wisconsin Breast Cancer")

#Box Plot: Radius Mean

boxplot(radius\_mean ~ diagnosis,data=wdbc,

horizontal=TRUE,

names=c("Benign","Malignant"),

col=c("green","red"),

xlab="Radius Mean",

main="Wisconsin Breast Cancer")