WISCONSIN BREAST CANCER DATA

JEFFERSON PARKER, PH.D.

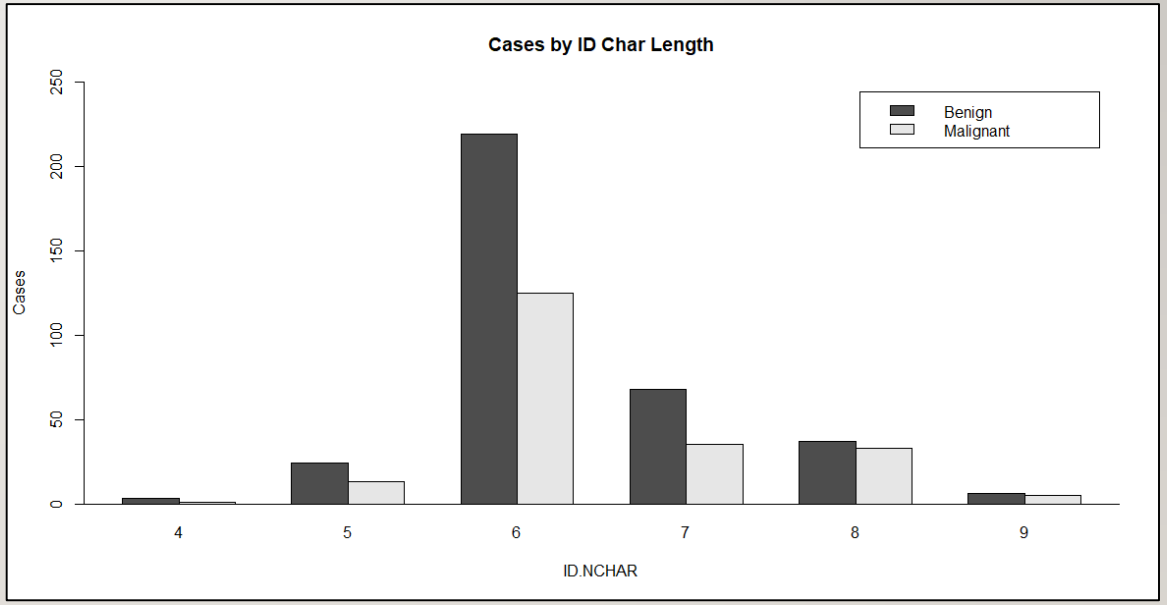
JANUARY 26, 2018

THE DATA

- Breast Cancer Wisconsin data set available from kaggle.com
  - Digitized micrographs from breast mass needle aspirate
  - Diagnosis as benign or malignant
  - 569 observations
  - 32 variables

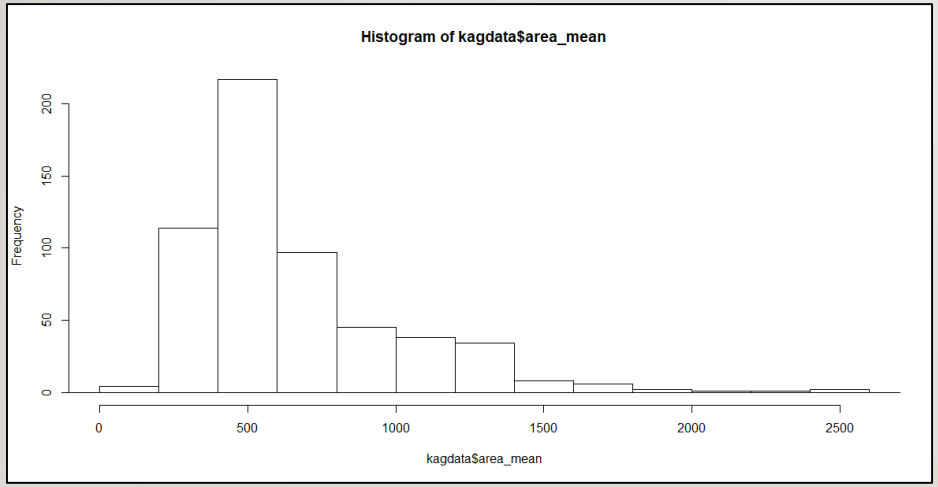
ID VARIABLE LENGTH

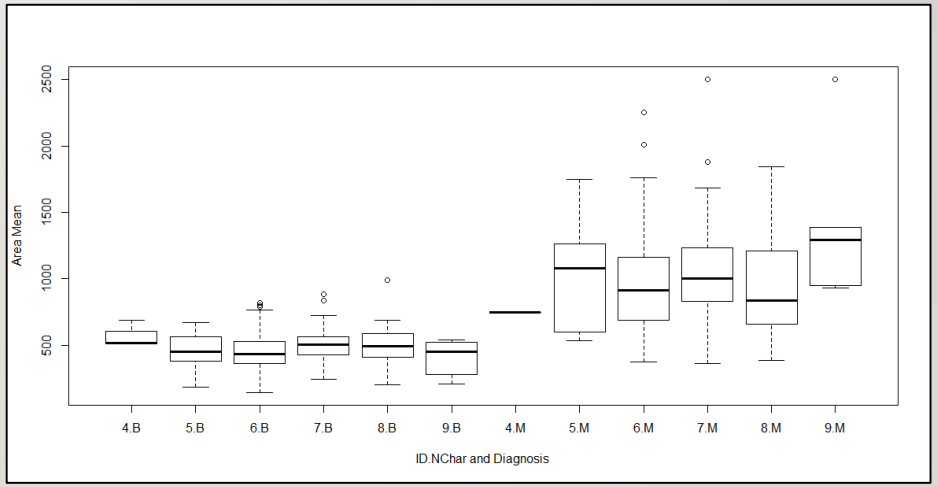
- The ID variable was a numeric value ranging from four to nine digits long.
  - The variety of ranges raised the question of whether the length was meaningful
  - Generated a new variable, ID.NCHAR to explore this further
- Large differences in samples per ID.NCHAR value



AREA_MEAN

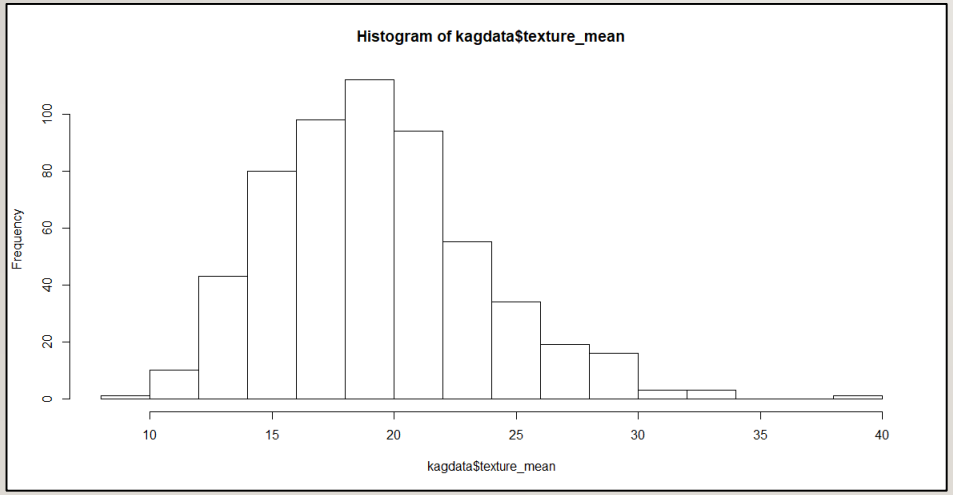
- The AREA_MEAN variable was defined as the *smoothness mean* of the image
- Not much difference among benign, more variation within malignant



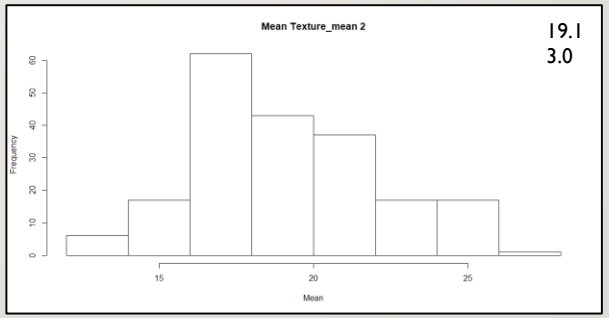


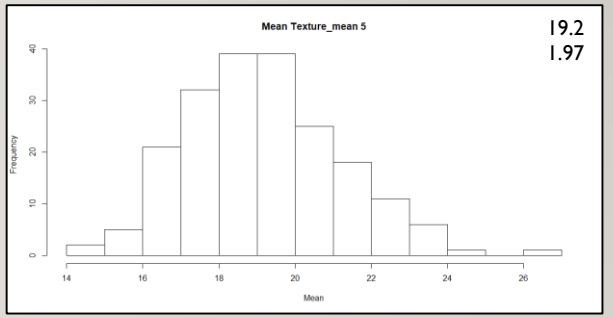
DATA DISTRIBUTION, CENTRAL LIMIT THEOREM TEXTURE_MEAN

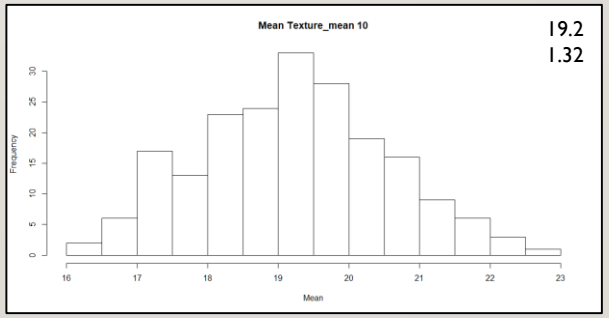
- The texture_mean variable measures the standard deviation of gray-scale values
  - Mean: 19.3
  - SD: 4.3

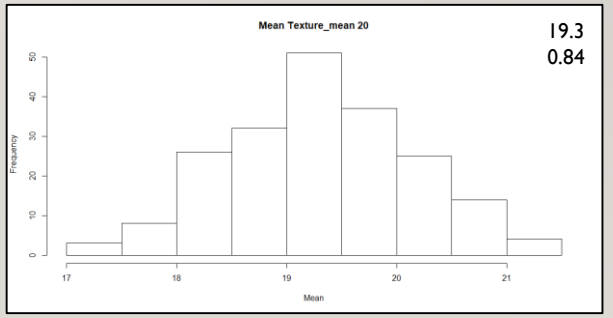


- Mean distribution of 200 samples of size 2, 5, 10 or 20









SAMPLING ID.NCHAR

- Two sampling methods were applied to the data to determine the impact on the ID.NCHAR variable in the sample group.
  - SRSWOR
  - Systematic

- Given the vastly different numbers of samples in each ID.NCHAR group, it is not surprising that some were not represented in the sample output

ID.Nchar	4	5	6	7	8	9
All	0.01	0.07	0.60	0.18	0.12	0.02
SRSWOR		0.10	0.75	0.15		
Systematic		0.1	0.7	0.1	0.05	0.05

QUESTIONS