Tutorial 8 Worksheet AY 22/23 Sem 1

DSA2101

The dataset for this tutorial comes from the UCI machine learning repository. The dataset `wdbc.data` can be obtained from Canvas or from here.

The dataset contains information on 569 cancer patients. We shall use it to explore the relationship between certain numerical features - features computed from images of the cancer cells, and whether the cancer is malignant or benign. Hence our response variable is *categorical*, while our features are all *numerical*.

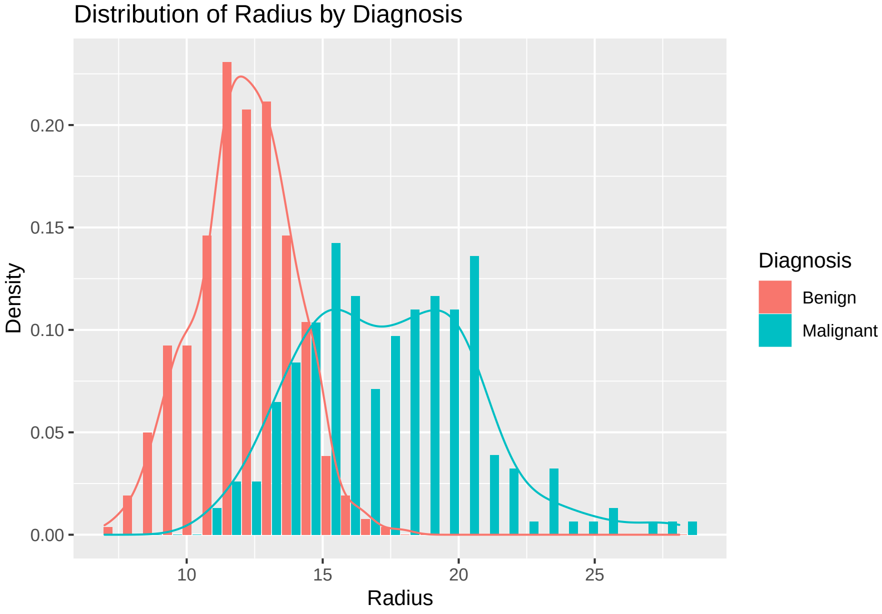
Here is a brief description of the columns in the data:

- Col. 1: ID number of patient
- Col. 2: Diagnosis of cancer type: M = malignant and B = benign.
- Col. 3 - 12: mean of ten different real-valued numerical features of cancer cells:
  - radius (mean of distances from center to points on the perimeter)
  - texture (standard deviation of gray-scale values)
  - perimeter
  - area
  - smoothness (local variation in radius lengths)
  - compactness (perimeter^2 / area - 1.0)
  - concavity (severity of concave portions of the contour)
  - concave points (number of concave portions of the contour)
  - symmetry
  - fractal dimension ("coastline approximation" - 1)
- Col. 13 - 22: resp. standard error of the above features.
- Col. 23 - 32: mean of the largest three of each of the above features.

We will only be working with the first 12 columns for this tutorial.

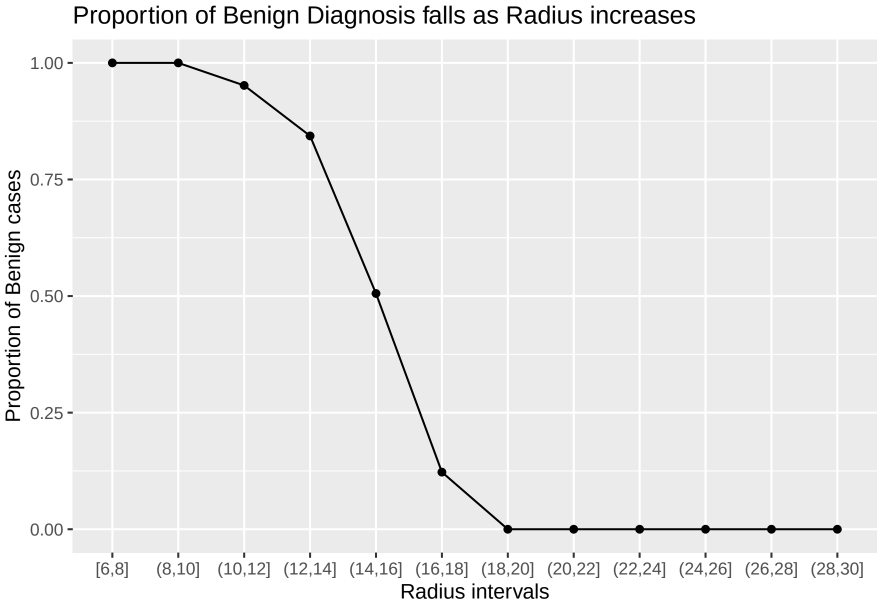
Histogram of radius

The following overlays the density plot of radius on the histogram, for each type of diagnosis. Recreate the plot as best you can.



Proportion of Benign Cancer

The plot above shows the distribution of radius for each type of diagnosis. However, when our response variable is categorical, we would be more interested in how the proportion of benign cancer changes as radius changes. Recreate this plot as best you can. It uses `cut_interval` to create intervals of the radius feature.



How can we update or improve this plot?

Data Exploration

Create a plot of your own using radius, diagnosis and one other numerical feature. How does this update your information of how diagnosis changes with radius?