STATS 501

Project Report

Team Members: Anmol Singh Suag

1. Introduction

1 in 8 women in the U.S. will develop symptoms of invasive breast cancer over the course of her lifetime [2]. A benign tumor is non-invasive and doesn't spread in the body, but on the other had, a malignant tumor is invasive and hence it's detection at early stages is critical. This project uses **UCI's Breast Cancer Wisconsin** dataset [1] and analyses the features of digitised images of breast mass for malignant and benign tumors. The project aims to understand the sample statistics and estimate the population parameters that would help researchers identify the class of tumor from digitised images. The project includes Hypothesis tests to verify the difference in concave points and symmetry among the 2 classes, as well as a regression analysis to find the association between radius and concave points.

2. Dataset

UCI Machine Learning Repository provides a dataset of 569 samples collected over time at University of Wisconsin[1]. For each sample, 10 features are computed from the digitised images of fine needle aspirate of breast mass. Each of the 569 samples is labeled as either **Benign (B) or Malignant (M)** based on the type of tumor identified. As the features are computed from images of a standardised size, most feature units are in pixels.

```
Benign Samples (B) = 357
Malignant Samples (M) = 212
Total = 569
```

2.1 Feature Description

A detailed description of the features is provided by Wolberg et al.[3]. Fig. 2.1.1 shows 2 records from the dataset.

	tumor_class	radius	texture	perimeter	area	smoothness	compactness	concavity	concave_points	symmetry	fractal_dimension
0	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871
1	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667

Fig. 2.1.1 Sample Dataset Rows

The 10 features provided for each of the 569 samples are as follows:

1. **Radius** : Mean of distances from center to points on the perimeter (pixels)
2. **Texture**: Standard deviation of gray-scale values in image (no unit)
3. **Perimeter** : The perimeter of cell (pixels)
4. **Area** : Area of cell in image calculated as number of pixels inside image (pixels*pixels)
5. **Smoothness**: It is the local variation in radius lengths (pixels)
6. **Compactness**: Calculated as (perimeter^2 / area - 1.0) (no units)
7. **Concavity**: Severity of concave portions of the contour (pixels)
8. **Concave points**: Number of concave portions of the contour (no units)
9. **Symmetry** : Symmetry is measured by finding the relative difference in length between pairs of line segments perpendicular to the major axis of the cell nucleus (pixels)

10. **Fractal dimension**: Coastline approximation - 1 (no units) [3]

2.2 Descriptive Statistics

First, the complete dataset is analysed and Table 2.2.1 shows the mean, median, max, standard deviation, 1st and 3rd Quartile of the individual features. Table 2.2.2 and Table 2.2.3 provide the same statistics for the features, but for malignant (M) and benign (B) samples respectively.

	radius	texture	perimeter	area	smoothness	compactness	concavity	concave_points	symmetry	fractal_dimension
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.048919	0.181162	0.062798
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.038803	0.027414	0.007060
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	0.049960
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.020310	0.161900	0.057700
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.033500	0.179200	0.061540
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.074000	0.195700	0.066120
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.201200	0.304000	0.097440

Table 2.2.1 Descriptive Statistics for all data

	radius	texture	perimeter	area	smoothness	compactness	concavity	concave_points	symmetry	fractal_dimension
count	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000
mean	12.146524	17.914762	78.075406	462.790196	0.092478	0.080085	0.046058	0.025717	0.174186	0.062867
std	1.780512	3.995125	11.807438	134.287118	0.013446	0.033750	0.043442	0.015909	0.024807	0.006747
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.000000	0.106000	0.051850
25%	11.080000	15.150000	70.870000	378.200000	0.083060	0.055620	0.020310	0.015020	0.158000	0.058530
50%	12.200000	17.390000	78.180000	458.400000	0.090760	0.075290	0.037090	0.023440	0.171400	0.061540
75%	13.370000	19.760000	86.100000	551.100000	0.100700	0.097550	0.059990	0.032510	0.189000	0.065760
max	17.850000	33.810000	114.600000	992.100000	0.163400	0.223900	0.410800	0.085340	0.274300	0.095750

Table 2.2.2 Descriptive Statistics Benign (B) samples

	radius	texture	perimeter	area	smoothness	compactness	concavity	concave_points	symmetry	fractal_dimension
count	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000	212.000000
mean	17.462830	21.604906	115.365377	978.376415	0.102898	0.145188	0.160775	0.087990	0.192909	0.062680
std	3.203971	3.779470	21.854653	367.937978	0.012608	0.053987	0.075019	0.034374	0.027638	0.007573
min	10.950000	10.380000	71.900000	361.600000	0.073710	0.046050	0.023980	0.020310	0.130800	0.049960
25%	15.075000	19.327500	98.745000	705.300000	0.094010	0.109600	0.109525	0.064620	0.174050	0.056598
50%	17.325000	21.460000	114.200000	932.000000	0.102200	0.132350	0.151350	0.086280	0.189900	0.061575
75%	19.590000	23.765000	129.925000	1203.750000	0.110925	0.172400	0.203050	0.103175	0.209850	0.067075
max	28.110000	39.280000	188.500000	2501.000000	0.144700	0.345400	0.426800	0.201200	0.304000	0.097440

Table 2.2.3 Descriptive Statistics Malignant (M) samples

Comparing Tables 2.2.2 and 2.2.3, we see that there is in fact a noticeable difference between the two classes of tumors! For example, the mean radius of malignant tumors (M) is 17.46 px whereas that of benign (B) tumors is 12.15 px, which is lesser. Similarly the mean area for malignant tumors (M) is 978.38 px^2 which is considerably higher than that of benign tumors with mean area of 462.79 px^2 . The fractal dimension mean and standard deviation of both the classes is nearly the same, which makes it this feature of little importance for classification.

2.3 Visual Analysis

Fig. 2.3.1 shows the box-plots of the attribute distributions drawn together for malignant (M) and benign (B) samples. As seen from Fig. 2.3.1, the distribution of concavity and texture of both the classes has many data points beyond the upper fence which are the outliers. The distribution of radius for the 2 classes has fewer outliers as compared to other features. Most of the box-plots are symmetric with mean being close to the median, and hence being approximately normally distributed i.e bell-shaped. Looking at the box-plot for distribution of concavity for benign (B) samples, it is seen that the upper whisker is longer, with a lot of outliers

beyond the upper fence. Hence, this distribution is a little right-skewed.

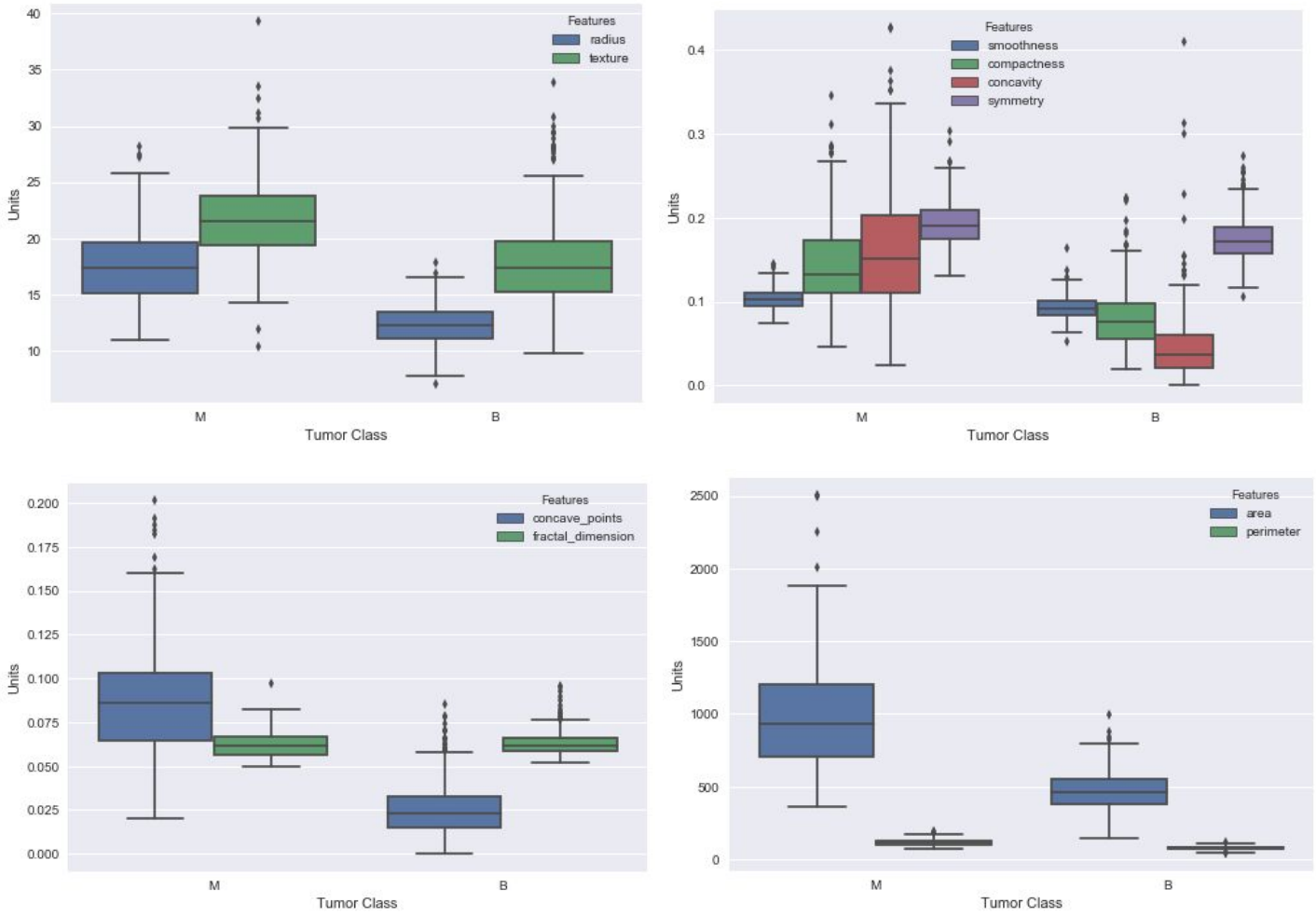

Fig 2.3.1 Comparing feature distribution Box-Plots for M and B

Fig. 2.3.2 shows the correlations among the 10 features for the collective data of the two classes. A high correlation between any two features is bad and one of the highly correlated feature is usually dropped for data modeling purposes.

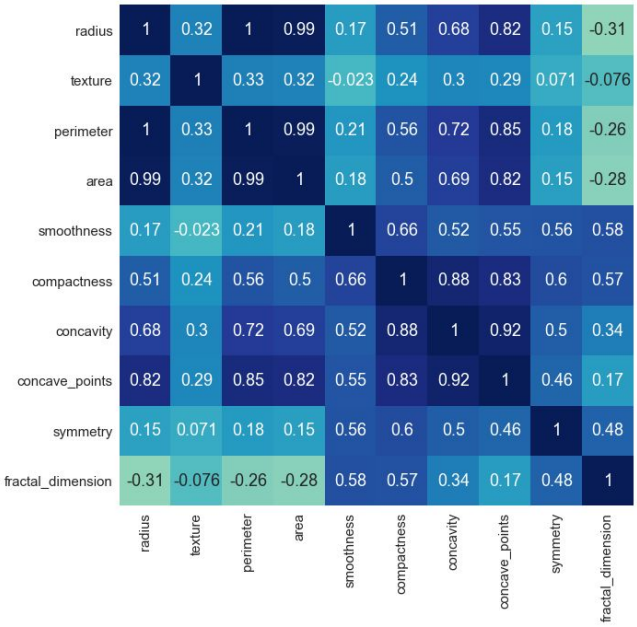

Figure 2.3.2 Correlation matrix for sample features

We see that radius is highly correlated with perimeter (1) and area (0.99), which makes sense as the 2-dimensional images of a tumor cell is like a disfigured circle with that radius. There seems to be a negative correlation between radius and fractal dimension (-0.31).

3. Data Analysis

We assume that the samples were drawn randomly and independently for both the classes. The sample size of class B, $n_B = 357$ and sample size of class M, $n_M = 212$. The samples sizes are large and hence by Central Limit Theorem, the sample mean for the features is approximately normally distributed, which is also visually confirmed as seen from Fig. 2.3.1. This section focuses on the radius, texture and concavity features of the class samples.

3.1 Confidence Interval for Population Mean

95% Confidence Intervals for population mean μ radius, texture and concavity for the malignant (M) and benign (B) tumors are calculated using the t-distribution as the population standard deviations are not known. We assume that the samples were drawn randomly and independently for both the classes and sample mean for the features is approximately normally distributed by CLT. Table 3.1 shows the 95% confidence intervals for population mean of radius, texture and concavity for the malignant (M) and benign (B) tumors.

Feature	**Benign (B)** $n_B = 357$	**Malignant (M)** $n_M = 212$
Radius (px)	$\bar{x} = 12.14,\ 95\%\ C.I\ of\ \mu = (11.96\ ,\ 12.33)$	$\bar{x} = 17.46,\ 95\%\ C.I\ of\ \mu = (17.03\ ,\ 17.90)$
Texture	$\bar{x} = 17.91,\ 95\%\ C.I\ of\ \mu = (17.50\ ,\ 18.33)$	$\bar{x} = 21.60,\ 95\%\ C.I\ of\ \mu = (21.09\ ,\ 22.12)$
Concavity (px)	$\bar{x} = 0.046,\ 95\%\ C.I\ of\ \mu = (0.042\ ,\ 0.051)$	$\bar{x} = 0.161,\ 95\%\ C.I\ of\ \mu = (0.151\ ,\ 0.171)$

3.2 Hypothesis Tests

1. A hypothesis test to verify if malignant (M) tumors **have higher concave points** that benign (B) tumors

The populations are assumed to be normally distributed and the sample is selected randomly and independently for both classes of tumor. The population variances of concave points are assumed to be equal as $s_M = 0.034$, $s_B = 0.016$ and $s_M^2/s_B^2 = 3 \times 10^{-7} < 3$. We perform a one-tailed 2 independent sample t-test.
$\alpha = 0.05$

$H_0 : \mu_M - \mu_B = 0$
$H_a : \mu_M - \mu_B > 0$

t-statistic= 29.35 with $n_M + n_B - 2 = 567$
p-value $\sim$ 0

As the p-value $\sim 0 < \alpha = 0.05$, we reject H_0 and accept H_a as the sample test provides us with enough evidence at this significance that the mean concave points in malignant tumor (M) is larger than that in benign tumor (B). We could have made a Type I error.

2. A hypothesis test to verify if malignant (M) and benign (B) tumors **don't have the same symmetry**

The populations are assumed to be normally distributed and the sample is selected randomly and independently for both classes of tumor. The population variances for symmetry are assumed to be equal as $s_M = 0.028$, $s_B = 0.025$ and $s_M^2 / s_B^2 = 5 \times 10^{-7} < 3$. We perform a two-tailed 2 independent sample t-test. $\alpha = 0.05$

$H_0 : \mu_M - \mu_B = 0$
$H_a : \mu_M - \mu_B \neq 0$

t-statistic= 8.34 with $n_M + n_B - 2 = 567$
p-value $\sim$ 0

As the p-value $\sim$ 0 < $\alpha = 0.05$, we reject H_0 and accept H_a as the sample test provides us with enough evidence at this significance that the mean symmetry in malignant tumor (M) is not the same as that in benign tumor (B). We could have made a Type I error.

3.3 Regression Analysis

The relationship between **cell radius and concave points** is found using linear regression. The complete dataset is used.
x : The independent variable, cell radius
y : The dependent variable, cell concave points

The Least Squares line of fit is found to be : $\hat{y} = -0.079 + 0.009x$
$a = -0.079$
$b = 0.009$
$R - Squared = 0.68$
$F - statistic = 1186$
$p - value = 4.35e - 141$

68% of the variability in concave points can be explained by the call radius. As the p-value is almost 0, the regression is a good fit. Fig. 3.3 shows the best-fit line in red slope $b = 0.009$ and intercept $a = -0.079$.

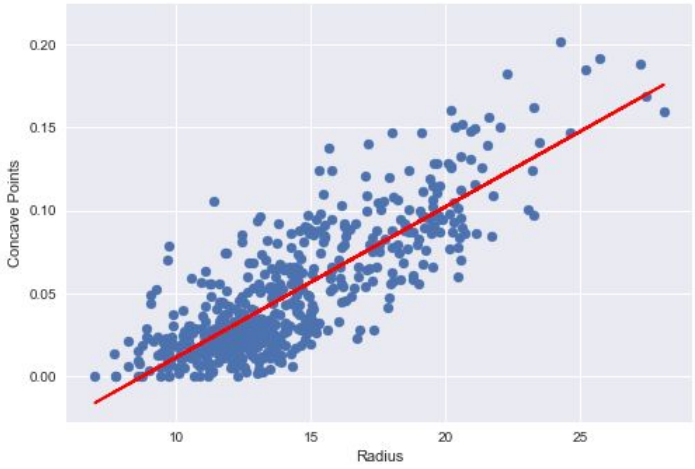

Fig 3.3 Regression Analysis

4. Conclusion

This project has helped visualise the distributions of various features like symmetry, radius, concavity of malignant (M) and benign (B) tumor cells. It was shown in section 3.1 with 95% Confidence intervals that the average radius, texture and concavity of malignant (M) cells is greater than that of benign (B) cells. In section 3.2 it was verified at 0.05 significance using Hypothesis tests that malignant (M) cells have higher concave points than benign (B) cells and that they have different symmetry . Section 3.3 calculates a best fit line to estimate cell concave points, given a radius.

The dataset comprised of 569 samples with 357 (B) and 212 (M) class samples. As the data was collected over the years at University of Wisconsin, it doesn't seem to be a good sample of global tumor cases. The samples could be biased towards a particular demography of individuals whose samples were taken to UWisc. This could have resulted in Type I errors in the Hypothesis tests in section 3.2. A better sampling method like, Stratified sampling could be used to have a good sample of demographic populations.

These statistical inferences could greatly help researchers differentiate the invasive malignant tumors from the benign tumors. The tests can be extended to find estimates to population parameters of malignant and benign tumors and help in their classification using machine learning algorithms or neural networks.

References

1. https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29
2. https://www.breastcancer.org/symptoms/understand_bc/statistics
3. https://dollar.biz.uiowa.edu/~street/research/cc97_01.pdf