**Spatial Difference Between Malignant and Benign Breast Cancer Cells**

**Math 421 - Applied Multivariate Statistics**

**Dr. Jahan**

**Lyndon Swarey**

**Nicholas Labella**

**August Horkan**

**Introduction**

Breast cancer is made up of left-over cells and are abnormal growths in the body. Depending on the type of cancer, it can grow and divide to create new cells that will replace dead cells. In general, benign breast cancer tends to be less severe and tends to not spread as rapidly as malignant breast cancer. Malignant breast cancer can grow at an exponential rate, causing severe tissue damage with special enzymes that dissolve the surrounding tissue and spreading to other parts of the body. Benign breast cancer will typically not contain any cancerous cells while malignant breast cancer contains cancerous cells. Benign breast cancer can still be painful and cause other serious problems. It also responds well if treated effectively and is much easier to remove. In some cases, the benign breast cancer is left untreated, and the person lives the rest of their life with the cancer. On the other hand, malignant breast cancer is often resistant to treatment and may require chemotherapy treatment and surgery to remove the cancer.

Since there are differences in how benign and malignant breast cancer behave and respond to treatment, there should also be differences in the size and shape of benign and malignant breast cancer cells. In this analysis, the Wisconsin breast cancer data set will be used to answer the question: is there a difference in average size and shape of benign and malignant breast cancer cells? The variables for this analysis will be identified using a combination of correlations and principal component analysis. The two sample Hotelling’s T2 test will be used to determine if there is a difference in the radius mean, smoothness mean and fractal dimension mean for benign and malignant breast cancer cells. Large sample simultaneous confidence intervals will also be used to determine which variables are statistically different for benign and malignant breast cancer cells.

**Data Description**

The Wisconsin breast cancer data set includes ID, 30 numerical variables and one categorical variable. The 30 numerical variables are mean, standard error and worst (the average of the three largest values) for radius (average distance from the center of the cell nucleus), texture (the standard deviation of gray scale values), perimeter, area, smoothness (the variation of lengths of the radius), compactness (perimeter^2/area-1), concavity (the severity of the concave portions of the cell’s contour), concave points (the number of concave points on the contour), symmetry and fractal dimension. These variables describe the cell nucleus in digital photos of fine needle aspirate of a breast cancer mass. The categorical variable is diagnosis. The diagnosis is malignant (M) or benign (B). There are 569 total observations with 357 of the observations classified as benign, and 212 of the observations classified as malignant. There are no missing values.

The standard error variables were removed from the analysis since standard error only tells us how much the sample mean varies from the population mean. It does not provide us with information about malignant and benign breast cancer. The worst variables or variables that are the average of the three largest values were also not used in the analysis. These variables have correlation of .700 or higher with their respective mean counterparts (Appendix, pg. 22). This implies that the mean variables and worst variables contain similar information. Fluctuations in the mean variable results in fluctuations in the worst variable. Principal component analysis was then used to determine which variables would be selected from radius mean, texture mean, perimeter mean, area mean, smoothness mean, compactness mean, concavity mean, concave points mean, symmetry mean, and fractal dimension mean.

After the principal component analysis had been run, principal component two was the most interesting since it showed a contrast between the size and shape of the cell nucleus (Appendix, pg. 23). The variables identified as size of the cell nucleus are radius mean and area mean since these variables have the largest negative eigenvalues. These variables have the following respective eigenvalues: -.314 and -.305. Radius mean was selected to be used in the analysis since radius mean has the largest eigenvalue, and radius mean and area mean have a correlation of .987 (Appendix, pg. 22). The high correlation between radius mean and area mean implies that both variables share similar information which makes sense since radius and area are related. For the shape variables, smoothness mean and fractal dimension mean were selected since they have the largest positive eigenvalues. Their eigenvalues are .402 and .572 respectively. In the analysis radius mean, smoothness mean and fractal dimension mean will be used to determine if there is a difference between the size and shape of benign and malignant breast cancer cells.

**Analysis**

In this analysis the variables radius mean, smoothness mean and fractal dimension mean are used to determine if there is a difference in the shape and size of benign and malignant breast cancer cells. To gain an understanding of whether there is a difference between benign and malignant breast cancer cells, it can be helpful to visually inspect the means.

|  |  |  |
| --- | --- | --- |
| **Variable Means by Breast Cancer Type** (Table 1) | | |
|  | **Benign** | **Malignant** |
| Radius mean | 12.1465 | 17.4628 |
| Smoothness mean | .0925 | .1029 |
| Fractal dimension mean | .0629 | .0627 |

Table 1 indicates that there is a difference in the means for benign and malignant breast cancer cells for radius mean and smoothness mean. Malignant breast cancer cells tend to have, on average, larger values for the radius mean and smoothness mean. On the hand, there does not appear to be much of a difference between benign and malignant breast cancer cells for fractal dimension mean. Fractal dimension mean tends to have a slightly smaller value for malignant breast cancer cells than for benign breast cancer cells. To determine if there is a difference between benign and malignant breast cancer cells, a two sample Hotelling’s T2 test will be used. This test will be used since we are wanting to test two mean vectors from two populations, benign breast cancer and malignant breast cancer, to determine if there is a difference in one or more of the variables. Large sample simultaneous confidence intervals will be used to determine which variables are statistically different.

To conduct a two sample Hotelling’s T2 test and large sample simultaneous confidence intervals, the assumptions need to be checked. The first assumption is that the observations for benign breast cancer were randomly selected from the population of people who have benign breast cancer and the observations for malignant breast cancer were randomly selected from the population of people who have malignant breast cancer. Since we were unable to find how the data was collected from each of the populations, we assume that the observations were randomly selected from their respective populations. This implies that the observations are independent of each other.

The second assumption is that both populations are multivariate normal. For benign breast cancer at the univariate levels, radius mean appears to be normally distributed. The qq-plot is in an approximate straight line. For smoothness mean, normality appears to be affected by an outlier. The p-values from the tests for normality are less than alpha = .05 suggesting that this variable may not be normally distributed. Fractal dimension mean also does not appear to be normally distributed. The qq-plot does not form an approximate straight line and the p-values from the tests for normality are all less than alpha = .05 suggesting that this variable may not be normally distributed (Appendix, pg. 14-17). At the bivariate level, the scatter plots of all pairwise combinations of radius mean, smoothness mean and fractal dimension mean are approximately elliptical (Appendix, pg. 18-19). This suggests radius mean, smoothness mean and fractal dimension mean may all be approximately bivariate normal. Even though the variables appear to be approximately bivariate normal, the assumption that the population is multivariate normal may not be met since the variables, smoothness mean and fractal dimension mean, are not approximately normal at the univariate level. According to the central limit theorem, the distribution is approximately multivariate normal for large sample sizes in which *n*, the number observations, is large in comparison to *p*, the number variables, regardless of the distribution of the underlying population. In this case, *n* = 357 and *p* = 3. Since *n* is significantly larger than *p*, we can assume that radius mean, smoothness mean and fractal dimension mean for benign breast cancer cells have an approximate multivariate normal distribution.

For malignant breast cancer at the univariate level, radius mean and smoothness mean are approximately normally distributed since the qq-plots for radius mean and smoothness mean are in an approximate straight line. On the other hand, fractal dimension mean does not appear to be normally distributed. Based on the qq-plot, there is an outlier that may affect normality. The p-values from the tests for normality are all less than alpha = .05. This suggests that fractal dimension mean may not be normally distributed for malignant breast cancer cells (Appendix, pg. 14-17). At the bivariate level, the scatter plots for radius mean and smoothness mean and for radius mean and fractal dimension mean are approximately elliptical suggesting that radius mean and smoothness mean and radius mean and fractal dimension mean are approximately bivariate normal. On the other hand, the scatterplot for smoothness mean and fractal dimension mean does not form an ellipse (Appendix, pg. 20-21). This suggests that smoothness mean and fractal dimension mean may not be bivariate normal. Thus, radius mean, smoothness mean and fractal dimension mean may not have a multivariate normal distribution. Since *n* = 212 is significantly larger than *p* = 3, the central limit theorem states that radius mean, smoothness mean and fractal dimension mean are approximately multivariate normal.According to the scatter plots and qq-plots, the assumption that both populations have an underlying multivariate normal distribution is not met. By using the central limit theorem, this assumption is met, and we can still proceed with the two sample Hotelling’s T2 test.

The third assumption is that benign and malignant breast cancer have the same covariance matrix. A test for homogeneity between the matrices was used to determine if the covariance matrix for benign breast cancer is equal to the covariance matrix for malignant breast cancer. This resulted in = 156.86 and a p-value of <.0001. Since the p-value is less than alpha = .05, there is evidence to suggest that the covariance matrix for benign breast cancer is not equal to the covariance matrix for malignant breast cancer. Thus, two sample Hotelling’s T2 test and large sample confidence intervals with unequal covariance will be used since the assumption that benign and malignant breast cancer have the same covariance matrix is not met.

The null and alternate hypothesis for the two sample Hotelling’s T2 test are as follows:

H0: The means of the radius mean, the smoothness mean, and the mean fractal dimension mean are the same for malignant and benign breast cancer cells.

H1: At least one of the means of the radius mean, the smoothness mean, and the fractal dimension mean are not the same for malignant and benign breast cancer cells.

The calculated T2 statistic was fairly large, at 646.47, which led to a much larger calculated F-value of 214.42 than the 95% critical F-value of 2.627. The p-value for the Hotelling's T2 test is not functionally different from zero (and was reported as zero using the computing software), so we reject the null hypothesis at alpha = .05. There is significant statistical evidence of a difference in means in at least one of the variables, radius mean, the smoothness mean, and the fractal dimension mean, between malignant and benign breast cancer cells.

|  |  |  |
| --- | --- | --- |
| **Simultaneous Confidence Intervals** (Table2) | | |
|  | **Lower** | **Upper** |
| Radius mean | 4.4323 | 6.1877 |
| Smoothness mean | .0066 | .0154 |
| Fractal dimension mean | -.0025 | .0026 |

The simultaneous 95% large sample confidence intervals (Table 2) of the difference in the means between malignant and benign breast cancer cells shows that the radius mean and the smoothness mean were significantly different from zero, while the difference in the fractal dimension mean between malignant and benign breast cancer cells was almost centered around zero, so there was no evidence of a difference for the fractal dimension mean. In other words, there is a statistical difference for radius mean and smoothness mean for benign and malignant breast cancer cells, but there is no difference between malignant and benign breast cancer cells for fractal dimension mean.

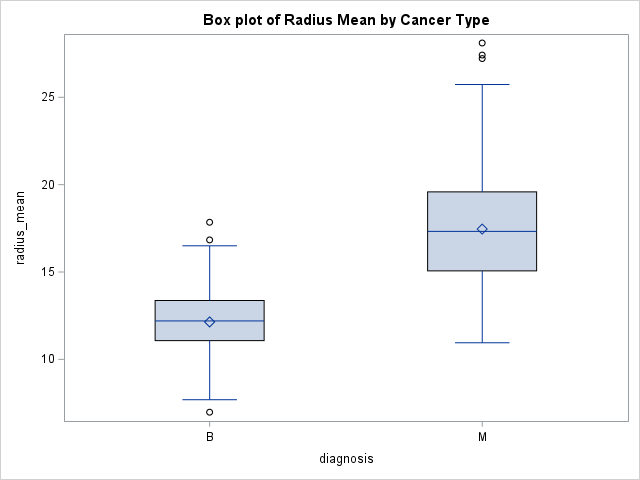
**Conclusion**

There are two primary and critical importance to there being statistical differences that medical equipment is able to read, such as radius, smoothness, and fractal dimension. One is for the medical staff that are looking at the data to be able to decide whether or not any critical procedures are necessary and which procedures may be more likely to cause harm than they are to lower the likelihood of harm, while only using the least invasive tests possible. Another is for allowing different data science methods to be able to classify an individual’s (or population’s) cancer status, to aid a medical staff or researchers to make critical decisions on what should be used or created in the pursuit of a better rate of survival and quality of life in the future.

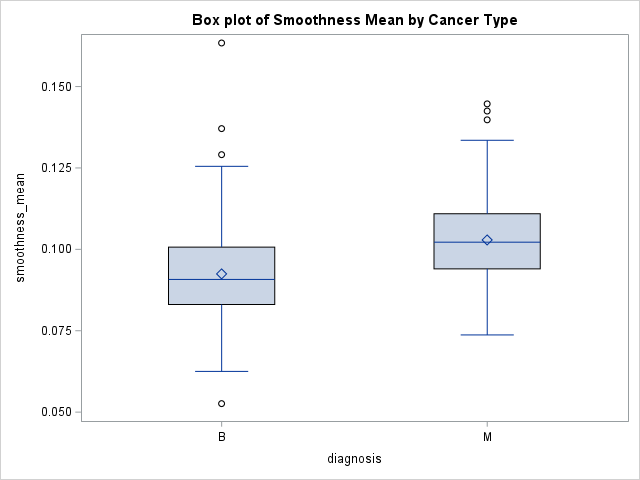
For these reasons, it is important that the radius mean and smoothness mean of a breast cancer cells are valuable in telling the difference between benign and malignant breast cancer cells. It is also very valuable to have a result that the mean fractal dimension does not have a statistically significant difference between benign and malignant breast cancer tumors. In particular, the fact that the test using all three variables had a p-value of essentially zero and an extremely high calculated F-value gives reason to believe that classification, at least to a certain extent, is very possible, even only given these three variables, although likely only needing the two variable which were found to be statistically significant, radius mean and smoothness mean.

**Appendix**

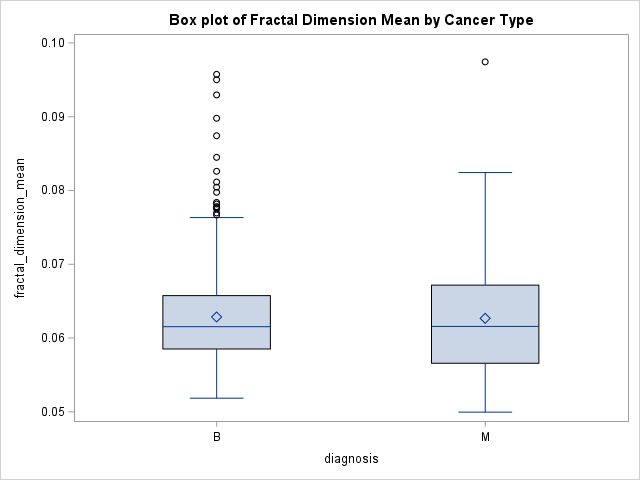
**Data Exploration**



Overall, malignant breast cancer (M) has a larger mean and median radius mean than benign breast cancer (B). Based on the boxplots, the radius mean for malignant breast cancer (M) is slightly right skewed while the radius mean for benign breast cancer (B) appears to be approximately normal. The radius mean for malignant breast cancer (M) has a larger range and interquartile range than benign breast cancer (B). For benign breast cancer (B), radius mean has three outliers with the values 6.981, 16.84 and 17.85, and the radius mean for malignant breast cancer M) has three outliers with the values 27.22, 27.42 and 28.11.

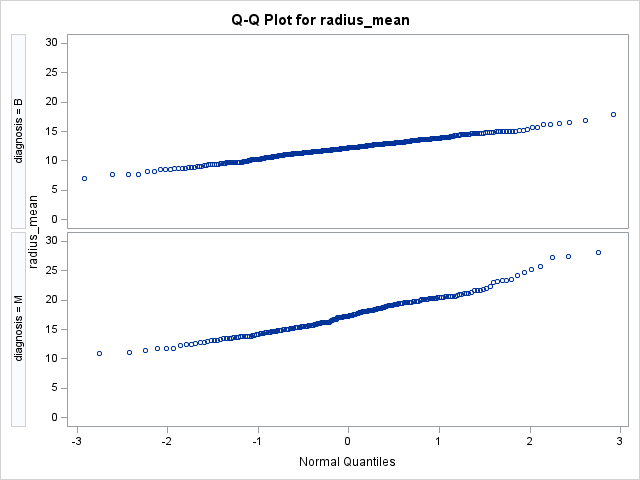


Since benign breast cancer (B) has a smaller mean and median in variation of radius length than malignant breast cancer (M), benign breast cancer (B) tends to be smoother than malignant breast cancer (M). Both benign (B) and malignant (M) breast cancer have similar interquartile ranges, but benign breast cancer (B) has a larger range than malignant breast cancer (M). Due to the outliers, benign breast cancer (B) is right skewed. On the other hand, malignant breast cancer (M) appears to be approximately normal. Benign breast cancer (B) has four outliers with the values .0526, .2591, .1371 and .1634 while malignant breast cancer (M) has three outliers with the values of .1398, .1425, and .1447.



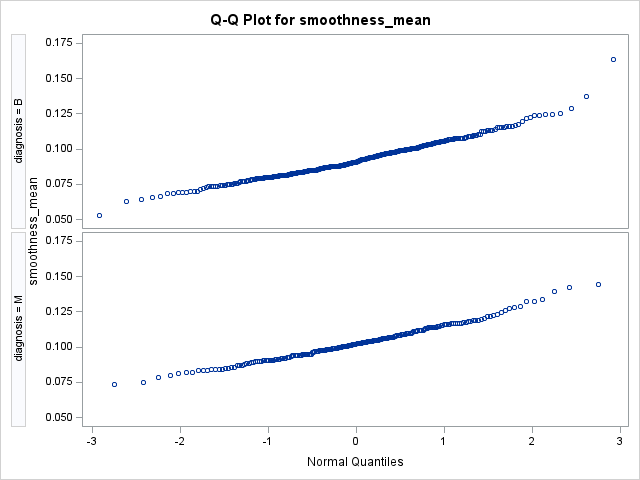
For fractal dimension mean, there does not appear to be a difference in the mean and median fractal dimension mean for benign (B) and malignant (M) breast cancer. Malignant breast cancer (M) has larger interquartile range and range than benign breast cancer (B). Due to outliers, benign (B) and malignant (M) breast cancer are right skewed. Malignant breast cancer (M) has an outlier with the value of .0974. Benign breast cancer (B) has 17 outliers with values greater than .0766.

**Test for Univariate Normal**



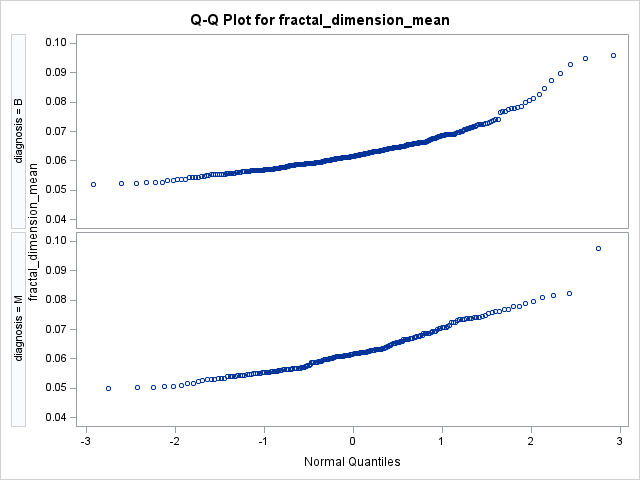
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test for Normality for Radius Mean** | | | | |
|  | **Benign (B)** | | **Malignant (M)** | |
|  | **Statistic** | **P-value** | **Statistic** | **P-value** |
| Shapiro-Wilk | W= .9967 | .6680 | W=.9777 | .0019 |
| Kolmogorov-Smirnov | D =.0328 | >.1500 | D=.0603 | .0597 |
| Cramer-von Mises | W-sq= .0539 | >.2500 | W-sq=.1167 | .0705 |
| Anderson-Darling | A-sq= .3878 | >.2500 | A-sq=.8843 | .0237 |

Since the points in the qq-plot for the radius mean for benign (B) and malignant (M) breast cancer are in an approximate straight line, there is evidence to suggest that variable radius mean is approximately normally distributed for both benign (B) and malignant (M) breast cancer. Radius mean for benign breast cancer (B) appears to have two possible outliers: one in the lower left corner with the value of 6.981 and one in the upper right corner with the value of 17.85. For malignant breast cancer, radius mean has four possible outliers: one in the lower left corner with the value of 10.95 and three in the upper right corner with the values of 28.11, 27.42 and 27.22. For both benign (B) and malignant (M) breast cancer, these outliers do not appear to have a significant effect on normality since the majority of the p-values are close to alpha =.05 or greater than alpha = .05.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test for Normality for Smoothness Mean** | | | | |
|  | **Benign (B)** | | **Malignant (M)** | |
|  | **Statistic** | **P-value** | **Statistic** | **P-value** |
| Shapiro-Wilk | W= .9755 | <.0001 | W=.9847 | .0215 |
| Kolmogorov-Smirnov | D =.0596 | <.0100 | D=.0388 | >.1500 |
| Cramer-von Mises | W-sq= .2062 | <.0050 | W-sq=.0777 | .2273 |
| Anderson-Darling | A-sq= 1.2656 | <.0050 | A-sq=.5693 | .1426 |

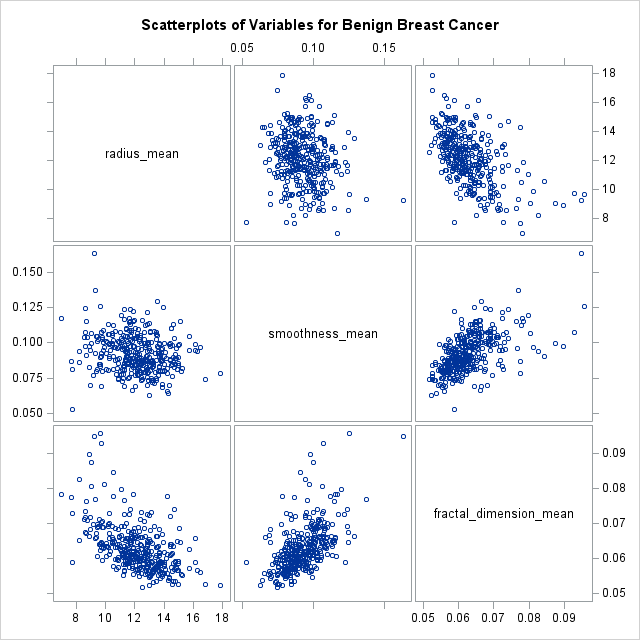
For benign breast cancer (B), smoothness mean has two outliers in the upper right corner with the value of .1371 and .1634. These outliers do appear to cause the points to curve up at the top right which could affect the normality for this variable since the p-values for the tests of normality are all less than alpha = .05. There is also an outlier in the lower left corner with the value of .0526. On the other hand, smoothness mean for malignant breast cancer (M) appears to be approximately normal since all the points are in an approximate straight line. There are four possible outliers: one in the lower left corner with value of .0737 and three in the upper right corner with the values of .1398, .1425 and .1447. These outliers do not appear to significantly affect normality for the smoothness mean for malignant breast cancer (M) since the majority of the p-values from the tests for normality are greater than alpha = .05



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test for Normality for Fractal Dimension Mean** | | | | |
|  | **Benign (B)** | | **Malignant (M)** | |
|  | **Statistic** | **P-value** | **Statistic** | **P-value** |
| Shapiro-Wilk | W= .8870 | <.0001 | W=.9508 | <.0001 |
| Kolmogorov-Smirnov | D =.0975 | <.0100 | D=.0947 | <.0100 |
| Cramer-von Mises | W-sq= 1.2313 | <.0050 | W-sq=.3777 | <.0050 |
| Anderson-Darling | A-sq= 7.7662 | <.0050 | A-sq=2.3024 | <.0050 |

Fractal dimension mean for benign breast cancer (B) is skewed to the right since the points curve up in the top right corner. There does appear to be two outliers: one in the lower left corner with the value of .0519 and one in the upper right corner with the value of .0958. The p-values from the tests for normality are also less than alpha = .05. Overall, fractal dimension for benign cancer (B) does not appear to be normally distributed. For malignant breast cancer, fractal dimension mean does have an outlier in the top right corner with the value of .0974 and one in the lower left corner with the value of .0500. Overall, the points appear to be in an approximate straight line, but the outlier in the top right corner may cause this variable to not be normally distributed since the p-values for the test for normality are all less than alpha = .05.

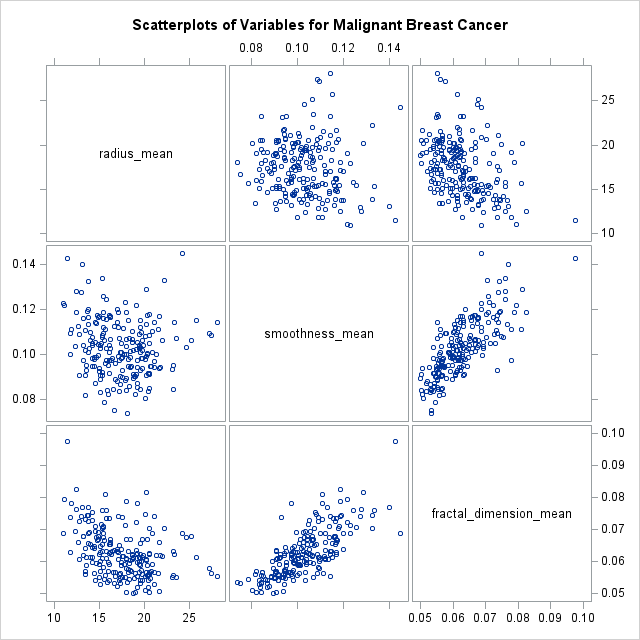
**Test for Bivariate Normal**



Based on the scatterplot of radius mean and smoothness mean for benign breast cancer, there is evidence that they are bivariate normal since the general shape of the scatterplot is in an approximate ellipse. There are five points that appear to be outliers since they are farther from the rest of the data. These points have the following coordinates assuming that radius mean is on the x-axis and smoothness mean is on the y-axis: (9.268, .1624), (6.981, .117), (7.76, .0526), (16.84, .0745) and (17.85, .0784). The scatter plot also indicates that there is little to no relationship between radius mean and smoothness mean.

For radius mean and fractal dimension mean for benign breast cancer, the general shape of the scatter plot is an approximate ellipse. This suggests that radius mean and fractal dimension mean are approximately bivariate normal. Assuming that radius mean is on the x-axis and fractal dimension mean is on the y-axis, radius mean and fractal dimension mean have a negative correlation. As radius mean increases, fractal dimension mean decreases. From the scatter plot, there appears to be eight potential outliers with the following coordinates assuming that radius mean is on the x-axis and fractal dimension mean is on the y-axis: (17.85, .0524), (16.84, .0527), (7.76, .0588), (9.04, .0874), (8.88, .0898), (9.73, .0930), (9.27, .0950) and (9.68, .0958). These points are farther from the rest of the points.

Since the scatter plot of smoothness mean and fractal dimension mean for benign breast cancer is an approximate ellipse, there is evidence that smoothness mean and fractal dimension mean are bivariate normal. Assuming that smoothness mean is on the x-axis and fractal dimension mean is on y-axis, the scatter plot also indicates that smoothness mean and fractal dimension mean have a positive correlation. As smoothness mean increases, fractal dimension mean increases as well. There are also seven possible outliers with the following coordinates assuming that smoothness mean is on the x-axis and fractal dimension mean is on y-axis: (.1634, .0950), (.1255, .0958), (.1072, .0930), (.0978, .0898), (.0997, .0874), (.1371, .0770) and (.0526, .0588). These points are not close to the other points in the scatter plot.



Based on the scatterplot of radius mean and smoothness mean malignant breast cancer, there is evidence that they are bivariate normal since the general shape of the scatterplot is in an approximate ellipse. There are five points that appear to be outliers since they are farther from the rest of the data. These points have the following coordinates assuming the x-axis is radius mean and the y-axis smoothness mean: (28.11, .0553), (27.42, .0562), (27.22, .0577), (20.18, .0814) and (11.42, .0974).. The scatter plot also indicates that there is little to no relationship between radius mean and smoothness mean.

For radius mean and fractal dimension mean for malignant breast cancer, the general shape of the scatter plot is an approximate ellipse. This suggests that radius mean and fractal dimension men are approximately bivariate normal. Assuming that radius mean is on the x-axis and fractal dimension mean is on the y-axis, radius mean and fractal dimension mean have a negative correlation. As radius mean increases, fractal dimension mean decreases. From the scatter plot, there appear to be six potential outliers with the following coordinates assuming that radius mean is on the x-axis and fractal dimension mean is on the y-axis: (17.85, .0524), (16.84, .0527), (7.76, .0588), (9.04, .0874), (8.88, .0898), (9.73, .0930), (9.27, .0950) and (9.68, .0958). These points are farther from the rest of the points.

Since the scatter plot of smoothness mean and fractal dimension mean for malignant breast cancer is not an approximate ellipse, there is evidence that smoothness mean and fractal dimension mean are not bivariate normal. Assuming that smoothness mean is on the x-axis and fractal dimension mean is on y-axis, the scatter plot also indicates that smoothness mean and fractal dimension mean have a positive correlation. As smoothness mean increases, fractal dimension mean increases as well. There are also five possible outliers with the following coordinates assuming that smoothness mean is on the x-axis and fractal dimension mean is on y-axis: (.1425, .0974), (.1398, .0769), (.1447, .0688), (.0974, .078) and (.0929, .0737). These points are not close to the other points in the scatter plot.

**Correlation Matrices**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations Between Mean and Worst Variables** | | | | | | | | | | |
|  | **Radius worst** | **Texture worst** | **Perimeter worst** | **Area worst** | **Smoothness worst** | **Compactness worst** | **Concavity worst** | **Concave points worst** | **Symmetry worst** | **Fractal dimension worst** |
| Radius mean | 0.970 | 0.297 | 0.965 | 0.941 | 0.120 | 0.413 | 0.527 | 0.744 | 0.164 | 0.007 |
| Texture mean | 0.353 | 0.912 | 0.358 | 0.344 | 0.078 | 0.278 | 0.301 | 0.295 | 0.105 | 0.119 |
| Perimeter mean | 0.969 | 0.303 | 0.970 | 0.942 | 0.151 | 0.456 | 0.564 | 0.771 | 0.189 | 0.051 |
| Area mean | 0.963 | 0.287 | 0.959 | 0.959 | 0.124 | 0.390 | 0.513 | 0.722 | 0.144 | 0.004 |
| Smoothness mean | 0.213 | 0.036 | 0.239 | 0.207 | 0.805 | 0.472 | 0.435 | 0.503 | 0.394 | 0.499 |
| Compactness mean | 0.535 | 0.248 | 0.590 | 0.510 | 0.566 | 0.866 | 0.816 | 0.816 | 0.510 | 0.687 |
| Concavity mean | 0.688 | 0.300 | 0.730 | 0.676 | 0.449 | 0.755 | 0.884 | 0.861 | 0.409 | 0.515 |
| Concave points mean | 0.830 | 0.293 | 0.856 | 0.810 | 0.453 | 0.667 | 0.752 | 0.910 | 0.376 | 0.369 |
| Symmetry mean | 0.186 | 0.091 | 0.219 | 0.177 | 0.427 | 0.473 | 0.434 | 0.430 | 0.700 | 0.438 |
| Fractal dimension mean | -0.254 | -0.051 | -0.205 | -0.232 | 0.505 | 0.459 | 0.346 | 0.175 | 0.334 | 0.767 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlations Between Mean Variables** | | | | | | | | | | |
|  | **Radius mean** | **Texture mean** | **Perimeter mean** | **Area mean** | **Smoothness mean** | **Compactness mean** | **Concavity mean** | **Concave points mean** | **Symmetry mean** | **Fractal dimension mean** |
| Radius mean | 1.000 | 0.324 | 0.998 | 0.987 | 0.171 | 0.506 | 0.677 | 0.823 | 0.148 | -0.312 |
| Texture mean | 0.324 | 1.000 | 0.330 | 0.321 | -0.023 | 0.237 | 0.302 | 0.293 | 0.071 | -0.076 |
| Perimeter mean | 0.998 | 0.330 | 1.000 | 0.987 | 0.207 | 0.557 | 0.716 | 0.851 | 0.183 | -0.261 |
| Area mean | 0.987 | 0.321 | 0.987 | 1.000 | 0.177 | 0.499 | 0.686 | 0.823 | 0.151 | -0.283 |
| Smoothness mean | 0.171 | -0.023 | 0.207 | 0.177 | 1.000 | 0.659 | 0.522 | 0.554 | 0.558 | 0.585 |
| Compactness mean | 0.506 | 0.237 | 0.557 | 0.499 | 0.659 | 1.000 | 0.883 | 0.831 | 0.603 | 0.565 |
| Concavity mean | 0.677 | 0.302 | 0.716 | 0.686 | 0.522 | 0.883 | 1.000 | 0.921 | 0.501 | 0.337 |
| Concave points mean | 0.823 | 0.293 | 0.851 | 0.823 | 0.554 | 0.831 | 0.921 | 1.000 | 0.463 | 0.167 |
| Symmetry mean | 0.148 | 0.071 | 0.183 | 0.151 | 0.558 | 0.603 | 0.501 | 0.463 | 1.000 | 0.480 |
| Fractal dimension mean | -0.312 | -0.076 | -0.261 | -0.283 | 0.585 | 0.565 | 0.337 | 0.167 | 0.480 | 1.000 |

**Eigenvalues from the Principal Component Analysis**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Eigenvalues from Principal Component Analysis** | | | | | | | | | | |
|  | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** | **Prin6** | **Prin7** | **Prin8** | **Prin9** | **Prin10** |
| Radius mean | 0.364 | -0.314 | -0.124 | 0.030 | 0.031 | 0.264 | -0.044 | 0.085 | 0.474 | 0.669 |
| Texture mean | 0.154 | -0.147 | 0.951 | 0.009 | 0.220 | 0.032 | 0.021 | -0.007 | 0.004 | 0.000 |
| Perimeter mean | 0.376 | -0.285 | -0.114 | 0.013 | 0.006 | 0.238 | -0.083 | 0.089 | 0.380 | -0.740 |
| Area mean | 0.364 | -0.305 | -0.123 | 0.013 | 0.019 | 0.332 | 0.261 | 0.145 | -0.747 | 0.032 |
| Smoothness mean | 0.232 | 0.402 | -0.167 | -0.108 | 0.844 | -0.062 | 0.011 | 0.171 | 0.006 | -0.004 |
| Compactness mean | 0.364 | 0.266 | 0.058 | -0.186 | -0.240 | -0.005 | -0.804 | 0.064 | -0.219 | 0.053 |
| Concavity mean | 0.396 | 0.104 | 0.041 | -0.167 | -0.313 | -0.601 | 0.367 | 0.450 | 0.081 | 0.010 |
| Concave points mean | 0.418 | 0.007 | -0.069 | -0.073 | 0.009 | -0.266 | 0.141 | -0.851 | -0.022 | 0.004 |
| Symmetry mean | 0.215 | 0.368 | 0.037 | 0.893 | -0.113 | 0.062 | 0.048 | 0.016 | 0.009 | -0.001 |
| Fractal dimension mean | 0.072 | 0.572 | 0.114 | -0.349 | -0.265 | 0.568 | 0.345 | -0.065 | 0.130 | -0.007 |

**References**

https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240

<https://www.healthline.com/health/cancer/difference-between-benign-and-malignant-tumors>

<http://pathology.jhu.edu/pc/BasicTypes1.php>

<https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html>

https://health.uconn.edu/orthopedics-sports-medicine/conditions-and-treatments/a-z-index/benign-malignant-tumor

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

**SAS Code**

**proc** **import** datafile="E:\data.csv"out=WORK.cancer(DROP =id radius\_se texture\_se

perimeter\_se area\_se smoothness\_se compactness\_se concavity\_se concave\_points\_se symmetry\_se

fractal\_dimension\_se)

dbms=csv replace;

getnames=yes;

**run**;

**PROC** **CORR** data= cancer noprob ;**run**;

**Data** cancer (keep=radius\_mean texture\_mean perimeter\_mean

area\_mean smoothness\_mean compactness\_mean concavity\_mean

concave\_points\_mean symmetry\_mean fractal\_dimension\_mean diagnosis);

set cancer;

**run**;

**proc** **means** data=cancer;

class diagnosis;

**run**;

**proc** **princomp** data= cancer out=cancer\_components;

**run**;

**proc** **discrim** data=cancer pool=test slpool=**.01** ;

class diagnosis;var radius\_mean smoothness\_mean fractal\_dimension\_mean;

**run**;

**PROC** **SORT** data = cancer out=cancer;

BY diagnosis;

**RUN**;

**PROC** **UNIVARIATE** DATA=cancer NORMAL;

class diagnosis;

Var radius\_mean smoothness\_mean fractal\_dimension\_mean;

QQPLOT;

**RUN**;

title "Scatterplots of Variables for Benign Breast Cancer";

**proc** **sgscatter** data=cancer;

where diagnosis = "B";

matrix radius\_mean smoothness\_mean fractal\_dimension\_mean;

**run**;

title "Scatterplots of Variables for Malignant Breast Cancer";

**proc** **sgscatter** data=cancer;

where diagnosis = "M";

matrix radius\_mean smoothness\_mean fractal\_dimension\_mean;

**run**;

title; **run**;

**PROC** **CORR** data = cancer NOMISS CSSCP NOPRINT NOPROB OUTP=T3S;

VAR radius\_mean smoothness\_mean fractal\_dimension\_mean;

BY diagnosis;

**RUN**;

**PROC** **PRINT** DATA=T3S;

**RUN**;

**PROC** **IML**;

START HTSQ3(N1,N2,P,V, alpha);

USE T3S WHERE(diagnosis="B");

READ ALL VAR V WHERE(\_TYPE\_='CSSCP')INTO A11;

READ ALL VAR V WHERE(\_TYPE\_='MEAN')INTO X1;

XBAR1 =X1`;

USE T3S WHERE(diagnosis="M");

READ ALL VAR V WHERE(\_TYPE\_='CSSCP')INTO A21;

READ ALL VAR V WHERE(\_TYPE\_='MEAN')INTO X2;

XBAR2 =T(X2);

A1= A11/(N1-**1**);

A2= A21/(N2-**1**);

ST=((**1**/N1)\*A1 + (**1**/N2)\*A2);

nu1 =**1**/N1\*(trace(((A1/N1)\*INV(ST))\*\***2**)+ (trace((A1/N1)\*inv(ST)))\*\***2**);

nu2 =**1**/N2\*(trace(((A2/N2)\*INV(ST))\*\***2**)+ (trace((A2/N2)\*inv(ST)))\*\***2**);

nu = (P+P\*\***2**)/( nu1 +nu2);

PRINT, A1, A2, ST, nu;

TSQ=(XBAR1-XBAR2)`\*INV(ST)\*(XBAR1-XBAR2);

Kp=INV(ST);

PRINT, XBAR1, XBAR2, Kp;

FCALC = ((nu -P +**1**)/(nu\*P))\*TSQ;

Pval=**1**-CDF('F',fcalc,P,nu -P +**1**);\* calculates P value;

fcrit=FINV(**1**-alpha,P, nu -P +**1**);\* calculates critical value of F;

PRINT, "T SQ = " TSQ, "FCALC = " FCALC, "P-value = " Pval, "F\_critical = " fcrit;

FINISH HTSQ3;

RESET NOPRINT;

V ={radius\_mean smoothness\_mean fractal\_dimension\_mean};

**RUN** HTSQ3(**357**,**212**,**3**,V, **0.05**);

**QUIT**;

**proc** **iml**;

q=quantile('CHISQ', **.95**, **3**);

Upper=**17.46**-**12.15**+((q)\*\***.5**)\*(**1.78**/(**357**\*\***.5**)+**3.20**/(**212**\*\***.5**));

Lower=**17.46**-**12.15**-((q)\*\***.5**)\*(**1.78**/(**357**\*\***.5**)+**3.20**/(**212**\*\***.5**));

radius\_mean = (Lower`) || (Upper`);

print radius\_mean[F=**8.4** C={"Lower" "Upper"}];

**QUIT**;

**proc** **iml**;

q=quantile('CHISQ', **.95**, **3**);

Upper=**.103**-**.092**+((q)\*\***.5**)\*(**.013**/(**357**\*\***.5**)+**.013**/(**212**\*\***.5**));

Lower=**.103**-**.092**-((q)\*\***.5**)\*(**.013**/(**357**\*\***.5**)+**.013**/(**212**\*\***.5**));

smoothness\_mean = (Lower`) || (Upper`);

print smoothness\_mean[F=**8.4** C={"Lower" "Upper"}];

**Quit**;

**proc** **iml**;

q=quantile('CHISQ', **.95**, **3**);

Upper=**.06287**-**.06268**+((q)\*\***.5**)\*(**.0067**/(**357**\*\***.5**)+**.0076**/(**212**\*\***.5**));

Lower=**.06268**-**.06268**-((q)\*\***.5**)\*(**.0067**/(**357**\*\***.5**)+**.0076**/(**212**\*\***.5**));

fractal\_dimension\_mean = (Lower`) || (Upper`);

print fractal\_dimension\_mean[F=**8.4** C={"Lower" "Upper"}];

**Quit**;

title "Box plot of Radius Mean by Cancer Type";

**proc** **sgplot** data=cancer;vbox radius\_mean / category=diagnosis; **run**;

title "Box plot of Smoothness Mean by Cancer Type";

**proc** **sgplot** data=cancer;vbox smoothness\_mean / category=diagnosis; **run**;

title "Box plot of Fractal Dimension Mean by Cancer Type";

**proc** **sgplot** data=cancer;vbox fractal\_dimension\_mean / category=diagnosis; **run;**