Summary of Results of Data Analysis : Doppler Ultrasound for Prediction of Malignant Thyroid Nodules

Koundinya Vajjha

1 Descriptive Statistics for Numeric Data

Here are the descriptive statistics for the numeric data.

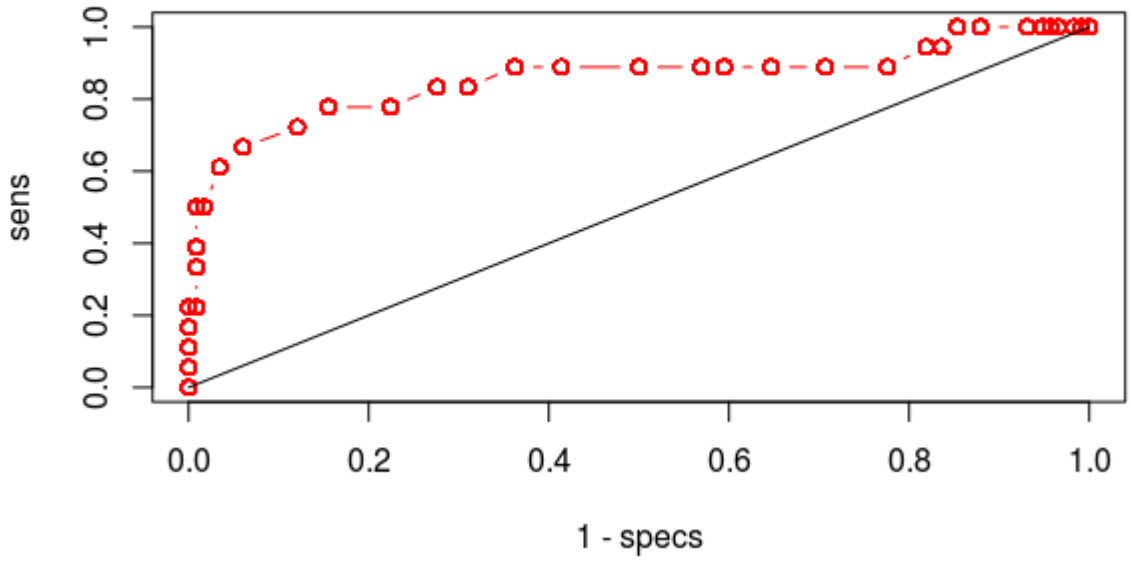
Table 1: Descriptive Statistics For Numeric Data

Statistic	N	Mean	St. Dev.	Min	Max
AGE	134	43.978	15.111	8	79
NO.OF.NODULES	134	1.440	0.710	1	3
AP	134	20.090	5.424	10	43
TRANS	134	21.567	4.995	13	40
VERTICAL	134	17.799	4.443	9	31
VASCULARITY	134	4.224	1.78398	1	9
RI	134	0.620	0.083	0.410	0.890

2 Sensitivity and Specificity Analysis.

Given below are the Sensitivity and Specificity for the variables DOPPLER.DIAG, VASCULARITY and Resistive Index (RI). The gold standard is taken to be the variable FINAL.DIAGNOSIS. The optimal cut-off for VASCULARITY has been set to be 4 as this gives the greatest Sensitivity and Specificity.

Finding the optimal cut-off for the variable RI is more complicated. First we vary the cutoff from 0 to 1 and find the sensitivity and specificity for each cutoff. Then we plot the Receiver Operating Characteristic (ROC) curve, which is 1-Specificity vs. Sensitivity. This is given below for the variable RI.



Then we find the point in the above graph which is closest to the point (0,1), which is called the "perfect classification" point. In our example, that point is (0.16,0.77). This point then gives our required sensitivity and specificity which is reported below.

Table 2:

Variable	Sensitivity	Specificity	Optimal cut-off
DOPPLER.DIAG	94%	97%	NA
VASCULARITY	72%	98%	≥ 4
RI	77%	84%	≥ 0.67

3 Association between variables and final diagnosis.

Given below are the crosstabulation tables and results of the Chi-square test for association between different variables and the final diagnosis. Henceforth,datanew will denote the dataset.

```
   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|
```

3.1 Variable : CAL

Total Observations in Table: 134

	datanew$FINAL.DIAGNOSIS		
datanew$CAL	BENIGN	MALIGNANT	Row Total
-	106	5	111
	1.022	6.587	
	0.955	0.045	0.828
	0.914	0.278	
	0.791	0.037	
COARSE	8	0	8
	0.167	1.075	
	1.000	0.000	0.060
	0.069	0.000	
	0.060	0.000	
FINE	1	12	13
	9.343	60.208	
	0.077	0.923	0.097
	0.009	0.667	
	0.007	0.090	
FINE+COARSE	1	1	2
	0.309	1.991	
	0.500	0.500	0.015
	0.009	0.056	
	0.007	0.007	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$CAL and datanew$FINAL.DIAGNOSIS
X-squared = 80.701, df = NA, p-value = 0.0004998
```

We reject the null hypothesis. So there is an association between CAL and FINAL.DIAGNOSIS.

3.2 Variable : SEX

datanew$SEX.INDICATOR	datanew$FINAL.DIAGNOSIS: BENIGN	MALIGNANT	Row Total
0	20	5	25
	0.125	0.803	
	0.800	0.200	0.187
	0.172	0.278	
	0.149	0.037	
1	96	13	109
	0.029	0.184	
	0.881	0.119	0.813
	0.828	0.722	
	0.716	0.097	
Column Total	116	18	134
	0.866	0.134	

```
 Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$SEX.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 1.1399, df = NA, p-value = 0.3273
```

Accept the null hypothesis. There is no association between SEX and FINAL.DIAGNOSIS.

3.3 Variable : No. of Nodules (1 and more than 1)

datanew$NODULE.INDICATOR	datanew$FINAL.DIAGNOSIS BENIGN	MALIGNANT	Row Total
0	34	8	42
	0.153	0.986	
	0.810	0.190	0.313
	0.293	0.444	
	0.254	0.060	
1	82	10	92
	0.070	0.450	
	0.891	0.109	0.687
	0.707	0.556	
	0.612	0.075	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$NODULE.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 1.6585, df = NA, p-value = 0.2629
```

Accept the null hypothesis. There is no association between No of nodules (1 and more than 1) and FINAL.DIAGNOSIS.

3.4 Variable : Echo genecity (hypo vs others)

	datanew$FINAL.DIAGNOSIS		
datanew$ECHO.INDICATOR	BENIGN	MALIGNANT	Row Total
0	79	5	84
	0.543	3.499	
	0.940	0.060	0.627
	0.681	0.278	
	0.590	0.037	
1	37	13	50
	0.912	5.879	
	0.740	0.260	0.373
	0.319	0.722	
	0.276	0.097	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$ECHO.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 10.833, df = NA, p-value = 0.0004998
```

Reject the null hypothesis. There is association between Echo genecity (hypo vs others) and FINAL.DIAGNOSIS.

3.5 Variable : Calcification (Fine vs others)

	datanew$FINAL.DIAGNOSIS		
datanew$CAL.INDICATOR	BENIGN	MALIGNANT	Row Total
0	115	6	121
	1.004	6.469	
	0.950	0.050	0.903
	0.991	0.333	
	0.858	0.045	
1	1	12	13
	9.343	60.208	
	0.077	0.923	0.097
	0.009	0.667	
	0.007	0.090	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$CAL.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 77.023, df = NA, p-value = 0.0004998
```

Reject the null hypothesis. There is association between Calcification (fine vs others) and FINAL.DIAGNOSIS.

3.6 Variable : Peripheral halo (absent vs others)

datanew$HALO.INDICATOR	datanew$FINAL.DIAGNOSIS BENIGN	MALIGNANT	Row Total
0	59	7	66
	0.061	0.393	
	0.894	0.106	0.493
	0.509	0.389	
	0.440	0.052	
1	57	11	68
	0.059	0.381	
	0.838	0.162	0.507
	0.491	0.611	
	0.425	0.082	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$HALO.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 0.89372, df = NA, p-value = 0.4528
```

Accept the null hypothesis. There is no association between Peripheral halo (absent vs others) and FINAL.DIAGNOSIS.

3.7 Variable : Vascularity (pattern more than 3 vs 3 or less)

datanew$VASCULARITY.INDICATOR	datanew$FINAL.DIAGNOSIS BENIGN	MALIGNANT	Row Total
0	114	5	119
	1.171	7.549	
	0.958	0.042	0.888
	0.983	0.278	
	0.851	0.037	
1	2	13	15
	9.293	59.889	
	0.133	0.867	0.112
	0.017	0.722	
	0.015	0.097	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$VASCULARITY.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 77.903, df = NA, p-value = 0.0004998
```

Reject the null hypothesis. There is association between Vascularity (pattern more than 3 vs 3 or less) and FINAL.DIAGNOSIS.

3.8 Variable : AP/Transverse ratio (> 1 vs other)

	datanew$FINAL.DIAGNOSIS		
datanew$AP.TRANS.INDICATOR	BENIGN	MALIGNANT	Row Total
0	100	7	107
	0.587	3.782	
	0.935	0.065	0.799
	0.862	0.389	
	0.746	0.052	
1	16	11	27
	2.326	14.989	
	0.593	0.407	0.201
	0.138	0.611	
	0.119	0.082	
Column Total	116	18	134
	0.866	0.134	

```
Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  datanew$AP.TRANS.INDICATOR and datanew$FINAL.DIAGNOSIS
X-squared = 21.684, df = NA, p-value = 0.0004998
```

Reject the null hypothesis. There is association between AP/Transverse ratio (> 1 vs other) vs FINAL.DIAGNOSIS.

4 Logistic Regression

To identify individual and combined predictors of malignancy (independent variables are: age,gender, no of nodules (1 and more than 1), echo genecity (hypo vs others), calcification (fine vs others), peripheral halo (absent vs others), vascularity (pattern more than 3 vs 3 or less), AP/Transverse ratio (> 1 vs other), resistive index).

We set up a logistic regression model. Since we have shown above that the variables Peripheral halo (absent vs others), No.of nodules (1 and more than 1) and Gender show no significant association with FINAL.DIAGNOSIS, we drop these variables and use the remaining for the logistic regression.

Here is the output.

Table 3: Logistic Regression Coefficients Along with Standard Error

	Dependent variable:
	FINAL.DIAGNOSIS.INDICATOR
AGE	-0.077
	(0.059)
ECHO.INDICATOR	-0.511
	(1.828)
CAL.INDICATOR	4.614^{*}
	(2.476)
VASCULARITY.INDICATOR	5.062^{***}
	(1.944)
AP.TRANS.INDICATOR	0.240
	(1.507)
RI	27.279^{***}
	(9.070)
Constant	-19.281^{***}
	(6.288)
Observations	134
Log Likelihood	-10.010
Akaike Inf. Crit.	34.021
Note:	*p<0.1; **p<0.05; ***p<0.01

In the above, the ECHO.INDICATOR is 1 if the data point is Hypo and 0 if it is anything else. Similarly for others. Here is the entire output.

```
Call:
glm(formula = FINAL.DIAGNOSIS.INDICATOR ~ AGE + ECHO.INDICATOR +
    CAL.INDICATOR + VASCULARITY.INDICATOR + AP.TRANS.INDICATOR +
    RI, family = "binomial", data = datanew, maxit = 100)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.08715  -0.11747  -0.05497  -0.01428   1.82815

Coefficients:
                      Estimate  Std. Error z value   Pr(>|z|)
(Intercept)          -19.28106    6.28834  -3.066    0.00217 **
AGE                   -0.07692    0.05940  -1.295    0.19532
ECHO.INDICATOR        -0.51072    1.82837  -0.279    0.77999
CAL.INDICATOR          4.61418    2.47576   1.864    0.06236 .
VASCULARITY.INDICATOR  5.06172    1.94382   2.604    0.00921 **
AP.TRANS.INDICATOR     0.24025    1.50699   0.159    0.87334
RI                    27.27857    9.07004   3.008    0.00263 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 105.735  on 133  degrees of freedom
Residual deviance:  20.021  on 127  degrees of freedom
AIC: 34.021

Number of Fisher Scoring iterations: 8
```