Michelle Zamperlini
Final Project – STAT 665

Association of BI-RADS score with Malignancy of Masses imaged in Breast Tissue

An exploration using bootstrapping tests of independence for ordinal variables when cell probabilities are near zero.

BACKGROUND

Mass biopsies account for a large portion of breast screening cost and overall patient anxiety, yet approximately 70% of biopsies return benign results. Using data collected by Dr. Schultz-Wendtland at the University Erlangen-Nuremberg's Institute of Radiology, this analysis explores the association between BI-RADS score and the severity (benign or malignant) of masses imaged in breast tissue (1).

BI-RADS (breast imaging and reporting data system) is a method of categorization used by radiologists to classify masses. The categories range from 1 to 5 representing the diagnostician's professional opinion on the likelihood that a mass is cancerous. Alongside the score, the system involves 3 attributes of the mass to be reported: margin, density, and shape. Each of these are also categorical. The goal of this exploration is to determine the rates of cancer present in samples from each BI-RADS category, there are other studies that analyze the attributes' individual predictive viability for determining breast cancer (2).

The following analysis draws inspiration from bootstraps methods of testing independence between ordinal variables (3) to test association between the ordinal variable, BI-RADS, and the response, severity. The methods in the paper showed the viability of the bootstrap approximations for overall small sample sizes, but it did not include unequal sample sizes or probabilities near the extremes of 0% or 100%. The uneven groups and ranging cell probabilities within this data make it a good case study for looking at the power of these bootstrap techniques under extreme circumstances.

DATA EXPLORATION

The data includes “BI-RADS” score, the outcome variable “severity”, a non-predictive variable “age”, and three categorical BI-RADS attributes “shape”, “margin”, and “density”. Our primary analysis focuses on the association of BI-RADS and severity.

	0	**Cat 2**	**Cat 3**	**Cat 4**	**Cat 5**	**6**	**mis**
Benign	2	13	30	427	40	3	1
Malignant	3	1	6	120	306	8	1

Figure 1: Table of frequencies for benign and malignant outcomes within each BI-RADS category.

BI-RADS is a system categorized to have 5 levels of severity. Initially exploration revealed one observation with a category of “55”, but this was deemed a clerical error and changed to a category of “5”. In **Figure 1**, you can see the frequencies of observations from the study. With only one mission value of BI-RADS per severity outcomes, these were deemed best to omit. For a 5 category system, though, a score of 0 and 6 does not make much sense for the analysis. A classification of 0 indicates the scan was inconclusive and the radiologists required further testing before any categorization could be done, meanwhile a classification of 6 means that there is prior knowledge of a mass and that mass is already shown to be malignant, likely resulting from a follow-up scan being requested (2).For the purposes of this analysis, both of these categories were dropped; 0 being a category of unknown BI-RAD classification and 6 being a category of prior knowledge.

With these values dropped, the final sample size is 938, with the majority of observations within categories 4 and 5, and category 2 with the smallest number of observations at 14. The cell probabilities by category, once we’ve cleaned the dataset, show exceedingly small values for Malignancy within Categories 2 and 3. Many statistics for association do not handle extreme values of 0% and 100% very well, and these cells can cause issues for analysis, we seek to see how the bootstrap method of independence behaves under these circumstances compared to standard methods of association.

	Benign	Malignant
Cat 2	0.013859275	0.001066098
Cat 3	0.031982942	0.006396588
Cat 4	0.455223881	0.126865672
Cat 5	0.042643923	0.321961620

Figure 2: A table of cell probabilities for outcomes by BI-RADS score

As for the binary variable “severity”, which includes levels 0 and 1 indication benign and malignant respectively, we will be treating it as an ordinal variable of 2 levels of increasing severity so as to utilize the bootstrap techniques for testing independence.

ANALYSIS

The methods described in Chan et al. involve selecting a statistic for testing the association, approximating the variance of that statistic using bootstrap resampling under the null hypothesis, and comparing to the traditional asymptotic testing procedures.

There are a few test statistics suitable for measuring association between ordinal variables, Goodman and Kruskal’s gamma (γ), Kendall’s tau-b (τ_b), Stuart’s tau-c (τ_c), and Somers’s D_{yx}, to name a few. Stuart’s tau-c adjusts for ties as well as table-size, making it generally the best option for situations where the variables being tested have different numbers of levels.

$$\tau_c = \frac{\Pi_a - \Pi_d}{1 - m^{-1}}, \qquad \tilde{\sigma}_0^2 = \frac{n}{(B-1)} \sum_{i=1}^{B} (\hat{\theta}_i^* - \bar{\theta}^*)^2$$

Figure 3: Formulas for Stuart’s tau-c and the bootstrap value for sigma

Figure 3 shows the formula for calculating Stuart’s tau-c and the subsequent variance we will be calculating according the bootstrap resampling method. The following steps for analysis were taken:

1. Convert the existing dataframe into a table of counts and use the *DescTools::StuartTauC()* to calculate the estimate of $\hat{\tau}_c = 0.63$.
2. Calculate the marginal probabilities of BI-RADS and Severity, then create a new resampling space with cell probabilities equal to the product of the marginals $\rho_{ij} = \rho_{i+}\,\rho_{+j}$.
3. Conduct bootstrapping with 1000 bootstrap samplings and calculate $\hat{\sigma}_0 = .39$ according to the formula in Figure 3.
4. At 95% confidence, compare the bootstrap method to the asymptotic test for Stuart's tau-c.

Method	$\hat{\tau}_c$	*Z-score/ Conf Interval*	*Reject at* $\alpha = .05$
Bootstrap	0.6323	Z = (19.648)	TRUE
Asymptotic	0.6323	Conf = (0.583, 0.682)	TRUE

Figure 4: Results from the bootstrap test and the asymptotic test for independence. StuartsTauC() output gives the confidence interval, but we can still establish rejection from this by whether or not 0 is included in the interval.

The results from the bootstrap approximation of the test statistic's variance and the asymptotic test both agree and reject the hull hypothesis of no association at the 95% confidence level. The estimate of $\hat{\tau}_c = 0.63$ indicates a positive association between the ordinal variables: an increase in BI-RADS Category is associated with an increase in Severity. This is precisely the association the variables hopes to achieve, indicating it is functioning as intended.

POWER

Despite the minimal sample sizes within Categories 2 and 3 compared to 4 and 5, both methods still detected an association and rejected the null. Perhaps this is because the link between Category 4 to 5 and severity are large enough to make the association known, or maybe the sample size is large enough that the small cell probabilities for Categories 2 and 3 are not disruptive to the association.

To test the two potentials, we conduct a power analysis similar to the analysis done in Chan et al., but with values specific to the targeted problem of the proportion of benign or malignant tumors within each BI-RADS category.

We test the methods on sample sizes n = (50, 100, 150, 300, 600) too see how well the methods do at detecting the association at smaller sample sizes. The following conditional probabilities for severity by BI-RADS score are generally accepted with studies into the system (4):

- BI-RADS 2 (benign): ~0% malignancy risk
- BI-RADS 3 (probably benign): ≤ 2% malignancy risk
- BI-RADS 4A (low suspicion): >2% to ≤10% malignancy risk
- BI-RADS 4B (moderate suspicion): >10% to ≤50% malignancy risk
- BI-RADS 4C (high suspicion): > 50% to < 95% malignancy risk
- BI-RADS 5 (probably malignant): ≥ 95% malignancy risk

The above probabilities will be used to establish the effect sizes for our power analysis, but our data set does not include the subcategories 4A, 4B, or 4C. Why that is the case is not entirely understood, it may be because the date the data was collected was prior to the latest BI-RADS procedure updates, perhaps it was not information given to the researchers, or it could be information lost in translation along the way of the dataset being shared. The effects at BI-RADS 2, 3, and 5 will be held constant at .05%, 2%, and 95% risk of malignancy respectively, but the probability of malignancy at BI-RADS category 4 will be tested at 5%, 10%, 50%, 90%, and 93%. If the question posed earlier is correct, and the association is strong enough between Categories 4 and 5 to make up for the small sample counts within Categories 2 and 3, if the probability of risk for Category 4 and 5 are much closer, then we may see the power of our methods fall. In all, the power analysis has 12,500 points of observation: 500 replications at 5 sample sizes for each of 5 underlying population conditional probabilities.

Since the study is retrospective the total sample size was the only fixed value, there was no pre-established quota for BI-RAD group sizes or proportion of malignancy cases. Therefore the conditional probabilities are converted into cell probabilities which are in turn sampled as a multinomial within r. The same analysis as prior is run on these

simulations and the rate of rejection is reported as a means to evaluate the power of these methods.

Sample Size = 50

Cat 4 risk	Asymptotic Rejection Rate	Bootstrap Rejection Rate
5%	1	1
10%	1	1
50%	.99	.99
90%	.41	.41
93%	.36	.36

Sample Size = 100

Cat 4 risk	Asymptotic Rejection Rate	Bootstrap Rejection Rate
5%	1	1
10%	1	1
50%	1	1
90%	.82	.82
93%	.65	.65

Sample Size = 150

Cat 4 risk	Asymptotic Rejection Rate	Bootstrap Rejection Rate
5%	1	1
10%	1	1
50%	1	1
90%	.89	.89
93%	.91	.91

Sample Size = 300

Cat 4 risk	Asymptotic Rejection Rate	Bootstrap Rejection Rate
5%	1	1
10%	1	1
50%	1	1
90%	1	1
93%	.99	.99

Figure 5: Tables that show the rejection rate of the asymptotic test and the bootstrap test. Each table corresponds to a different sample size.

The asymptotic and bootstrap tests performed the exact same for each level of risk for category 4 as well as across the sample sizes. It seems that for this type of disparity in group sizes and probability differences, both tests are equally powerful. The biggest issue occurred in that the tau-c estimation would return NA, which would cause both tests to fail to run. This occurs when the cell probabilities are too small in the groups. As the probability for Category 4 reaches similar probabilities for Category 5, though, we see a small decrease in the rate of rejection, meaning it's more difficult to attribute the association.

SUPPLEMENTAL ANALYSIS

The primary purpose of this analysis was to look at the association between BI-RADS categories and severity, by doing so we begin to get an idea about whether the biopsies being conducted are truly necessary. The analysis established an association between BI-RADS and malignancy in those who were biopsied, but in the following short supplemental analysis we look at the additional attributes of BI-RADS. If the attributes of BI-RADS are more predictive of malignancy than the BI-RADS categories themselves, then it's possible that biopsy decisions based solely on BI-RADS categories is leading to more biopsies than necessary.

To analyze this, we fit two logistic regression models. The first, $\text{SEVERITY} \sim \text{BI-RADS} + \text{AGE}$, looks at the relationship between BI-RADS and severity, controlling for the non-predictive integer variable "age". Age is a known risk factor for cancer diagnosis, but BI-RADS specifically lists age as a non-predictive indicator for BI-

RADS, meaning the association between severity and these ordinal categories should apply no matter the age group, therefore we include it in the model to control for its influence.

The second model fit is $\text{SEVERITY} \sim \text{MARGIN} + \text{SHAPE} + \text{DENSITY} + \text{AGE}$. Here we look at how predictive the attributes of BI-RADS are of malignancy. The results are summarized in Figure 6. Since age is included in both models, and not predictive, when listing the models from hence forward, this variable will not be written, but it is assumed included and was part of the analysis.

Model	*df*	*Residual Deviance*	*AIC*	*BIC*
BI-RADS	811	656.73	666.73	690.25
Density + Margin + Shape	804	714.26	738.26	794.72

Figure 6: The results of logistic regression models using the predictive variables and severity as the outcome.

The predictive power of age is so large that it is overshadowing the predictive significance of the other predictors, but this is fine, as we are looking at the overall model's deviance as a method of comparing their fit to severity, not conducting variable selection. The results above, comparing on residual deviance, AIC, and BIC, all indicate that BI-RADS is a better fit for the log odds of a mass being malignant. This demonstrates the value of professional expertise in evaluating the individual attributes that make up the BI-RADS system, as the attributes by themselves were not as good of a fit.

The initial research question, though, inquired about unnecessary biopsies. These would stem from a prediction model that has a lower specificity. When it comes to health, it's customary to take the mindset of "better safe than sorry", therefore researchers prioritize higher sensitivity (ability to detect true cases as true), which has a natural trade off with specificity. The ROC plots below show that for a sensitivity below approximately .8, the BI-RADS model had a better specificity, which would result in fewer unnecessary biopsies for that given level. For a sensitivity above approximately .8, though, the attributes have a higher specificity. Without expertise in this domain, I cannot say which model is better for the situation, it truly depends on how highly the experts in this field value sensitivity over specificity.

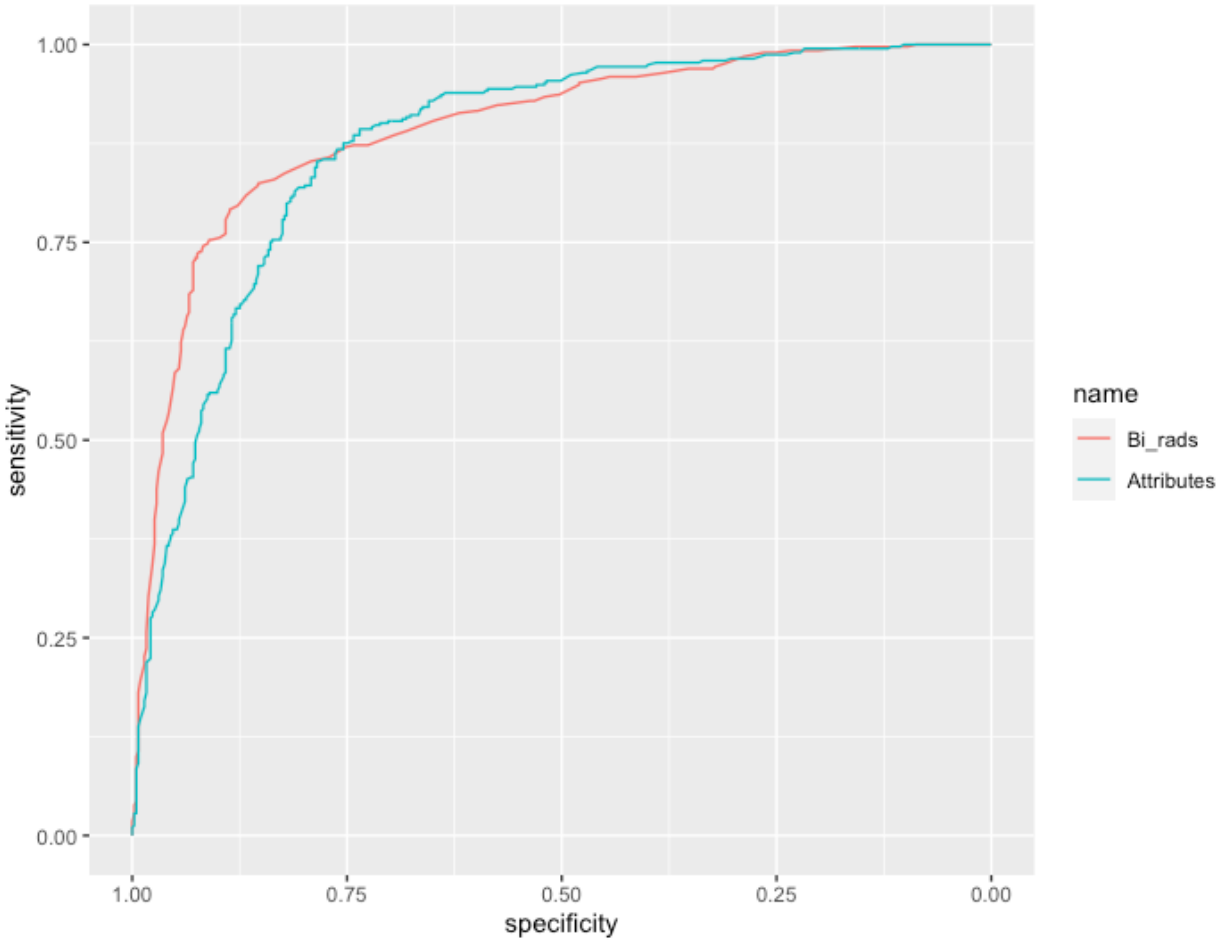

Figure 7: ROC plots of SEVERITY ~ BI-RADS + AGE, and SEVERITY ~ SHAPE + MARGIN + DENSITY.

REFERENCES

(1) Mammographic Mass Data. Elter. UC Irvine Machine Learning Repository. DOI: 10.24432/C53K6Z

(2) The Mammographic Density of a Mass Is a Significant Predictor of Breast Cancer. Woods et al. NIH Pubmed. DOI: 10.1148/radiol.10100328

(3) Chan W, Yung YF, Bentler PM, Tang ML. Tests of independence for ordinal data using bootstrap. Educational and Psychological Measurement. 1998;58(2):221-240. doi:10.1177/0013164498058002006

(4) Naser Ghaemian et al. Accuracy of mammography and ultrasonography and their BI-RADS in detection of breast malignancy. NIH Pubmed. doi: 10.22088/cjim.12.4.573