Reclassifying inferential statistics into diagnostic and predictive statistics with an application on gynecologic cancer

Abstract
Statisticians use to classify Statistics into two main parts, namely Descriptive and Inferential Statistics. Here, we suggest reclassifying Inferential Statistics into two parts, namely Diagnostic Statistics and Predictive Statistics. Based on that we will have four levels to analyze data (Descriptive, Diagnostic, Predictive and Perspective Statistics). Descriptive statistics mainly related to Graphs, Frequency tables, Measures of Central Tendency, Measures of Variation and Measures of Shape. Diagnostic statistics mainly related to the effects of the Independent variables (inputs) on the Dependent (Target) variable based on the Tests of Correlation or Association, Tests for Means differences and Tests for Classification. Predictive statistics mainly related to Estimation, Regression techniques and Time series Analysis for the Dependent (Target) variable. Perspective statistics mainly related to the previous three levels and acts as a prescription to how to solve or prevent the problem. In this paper, we will clarify the statistical tests used in each level of statistical analysis and will give an example on a real data related to Gynecologic Cancer

Keywords: inferential statistics, diagnostic statistics, predictive statistics, perspective statistics, gynecologic cancer

Introduction
Statisticians use to classify Statistics into two main parts, namely Descriptive and Inferential Statistics. Here, we suggest reclassifying Inferential Statistics into two parts, namely Diagnostic Statistics and Predictive Statistics. The Diagnostic statistics depends on Tests of Differences and Associations, while the Predictive statistics depends on Estimation, Prediction and Forecasting as shown in Figure 1.

We will consider having four levels of statistical analysis, namely Descriptive, Diagnostic, Predictive and Perspective statistics and will summarize the statistical tools that should be used. In terms of complexity of the algorithms and techniques involved, descriptive analytics are the simplest. Both Descriptive and Diagnostic statistics are related to the data already collected, and hence considered to be related to “past”. While Predictive and Perspective statistics are related to what is expected to happen, and hence considered to be related to “future”. Prescriptive analytics has the most impact on decision making, as it helps to identify the best action for the future. Predictive statistics mainly related to the previous three levels and acts as a prescription to how to solve or prevent the problem. Figure 2 shows a relative comparison of the four different types of analytics related to time.

Descriptive Analytics
Descriptive analysis is the statistical tools that should answer the question “What had happened?”. This form of analytics mainly deals with understanding the already gathered data. It is mainly related to Graphs, Frequency tables, Measures of Central Tendency, Measures of Variation and Measures of Shape. It involves the use of tools and algorithms to understand the internal structure of the Data and find categorical or temporal patterns or trends in it. The statistical tools may be summarized in the Table 1, based on the type of variable and its measurement level.
Reclassifying inferential statistics into diagnostic and predictive statistics with an application on gynecologic cancer

Table 1 Basic statistical tools for descriptive statistics

| Qualitative | Quantitative Interval or Ratio |
|-------------|--------------------------------|
| **Nominal** | **Ordinal** | **Bar, Pie** | **Bar, Pie** | **Bar** (for discrete), **Histogram, Polygon, Curve, Ogive** (for continuous), **Line** (for time), **Scatter diagram**, (for binary data). |
| **Tendency** | **Mode** | **Mode, Median** | **Mode, Median, Mean, Geometric mean, Harmonic mean, trimmed mean.** |
| **Variation** | **Quartile range** | **Range, Variance, Standard deviation, Coefficient of variation** |
| **Position** | **Quartiles** | **Standard Scores, Percentiles, Quartiles and Deciles, Skewness, Kurtosis** |

**Diagnostic Analytics**

Once the data is described, the next step is to seek independent variables affecting the Target (Dependent) variable, through answering the question “Why did it happened?”. Diagnostic analytics focuses on the reasons behind the observed patterns that are derived from descriptive analytics. The principal point here is the Target variable’s measurement level and its relation with the Independent variables (inputs). This may be checked mainly through the Tests for Means values for the Target using tests of differences and the Tests of Association for the Target with the inputs. Based on the Target’s measurement level, the statistical tools related to the differences, (where the inputs’ are categorical) may be summarized in Tables 2, while the statistical tools related to the association (for any Input’s measurement level) may be summarized in Table 3.

Table 2 Basic Diagnostic statistics tools for checking the differences in the target

| Dependent | Target (Dependent) Measurement level | Quantitative Interval or Ratio Scale (from Normal Population) |
|-----------|------------------------------------|----------------------------------------------------------|
| **Independent** | **Qualitative** | **Ordinal** (Rank) | **Scale (from Non-Normal Population)** |
| 2 independent groups | Chi-square test | Mann-Whitney | Mann-Whitney |
| 3+ independent groups | Chi-square test | Kruskal-Wallis | Kruskal-Wallis |
| 2 matched groups | McNemar | Wilcoxon test | Wilcoxon test |
| 3+ matched groups | Chi-square test | Friedman test | Friedman test |

Table 3 Basic diagnostic statistics tools for checking the association with the target

| Dependent | Target (Dependent) Measurement level | Quantitative Interval or Ratio |
|-----------|------------------------------------|-------------------------------|
| **Independent** | **Qualitative** | **Ordinal (Rank)** | **Interval or Ratio** |
| 2 independent groups | Phi | Contingency Coefficient Lambda | Bi-serial | Eta |
| Bi-serial | Kendall tau | Spearman Rho Gamma | Ordinal Bi-serial |
| Point Bi-serial Eta | Ordinal Bi-serial Pearson |
Predictive Analytics

Given the current trends in data identified by both the descriptive and diagnostic analytics, what might happen in the future is a crucial question. Predictive analytics tools provide insights into the possible future scenarios. Predictive analytics uses the outcomes of descriptive and diagnostic analytics to create a model for the future. In other words, analyzing what happened and gives insights to prepare a model for what is possible in the future, through answering the question “Why is likely to happen?”. The principal point here, as in the Diagnostic analytics, is the Target variable’s measurement level and its relation with the Independent variables (inputs). This may be checked mainly through the Estimation (in case of unknown population parameter), or Prediction (mainly based on regression techniques) or Forecasting future value (based on time series techniques), as indicated in Figure 1 and summarized in Table 4.

Table 4 Basic predictive statistics tools for checking the association with the target

| Dependent | Objective | Target (Dependent) Measurement level | Quantitative |
|-----------|-----------|-------------------------------------|--------------|
|           | Estimation| Nominal | Ordinal (Rank) | Interval or Ratio |
|           | Prediction| Logistic Regression, Generalized Linear Mixed Model | Ordinal Regression Generalized Linear Model, Generalized Linear Mixed Model | Linear, Non linear, General linear model, Generalized Linear Model, Generalized Linear Mixed Model |
|           | Forecasting| NA | NA | Exponential Smoothing, ARMA, ARIMA, SARIMA |

Prescriptive Analytics

Prescriptive analytics tools provide a “what if” kind of analysis capability. What are the different options available and which among them is the best suited, given the predictions and other constraints, through answering the question “What should happen?”. It is a result of Diagnostic and Predictive analytics. Through Prescriptive analytics, we advise an action or a solution to be taken before the occurrence of the problem. Simulation, Decision Modelling and Expert systems play the main role with Prescriptive statistics.

An application on Gynecologic Cancer

Gynecologic cancer (malignant tumor) is any cancer that starts in a woman’s reproductive organs. Types of Gynecologic Cancer is: Cervical cancer, Ovarian cancer, Uterine cancer (Uterine cancers can be one of two types: endometrial cancer (common) and uterine sarcoma (rare)), Vaginal cancer and Vulvar cancer. Each Gynecologic cancer is unique, with different signs and symptoms, different risk factors (things that may increase your chance of getting a disease), and different prevention strategies. All women are at risk for Gynecologic cancers, and risk increases with age. When Gynecologic cancers are found early, treatment is most effective. The following analysis are based on data collected from King Abdulaziz University Hospital, Saudi Arabia, during the period from beginning of the year 2000 to the end of 2016.

Descriptive analytics

We have initial routine tests of 513 patients (228 Benign 285 Malignant), with 118 fields: [Age, Nationality, Body Mass Index (BMI), Parity, Miscarriage, Date of admission, Marital status (Married before or not), Medical illness (Contains 60 types of illness), Previous surgery (Contains 47 type of surgery) and Heart block (HB). Date of admission start from 2000-01-16 to 2016-11-09. Table 5 summarize the highest frequencies for the patients’ diseases or previous surgery, while table 6 summarize the descriptive statistics for the continuous variables.
Table 6 Descriptive statistics for the continuous variables

| Variable  | Min | Max | Mean  | SD    | Skewness |
|-----------|-----|-----|-------|-------|----------|
| AGE       | 13  | 95  | 51.9  | 11.693| 0.468    |
| Parity    | 0   | 13  | 4.4   | 3.356 | 0.418    |
| Miscarriage| 0  | 8   | 0.5   | 1.099 | 2.761    |
| BMI       | 14.6| 168.4| 31.3  | 9.808 | 6.107    |
| HB        | 6.8 | 18.7| 11.4  | 1.741 | -0.006   |

Diagnostic analytics

Among all the variables measured, only Age, Parity and Miscarriage are the only significant continuous variables affecting the dependent variable (Tumor), this was done through using the independent t-test, with p-values less than 0.05, as shown in table 7. While Previous marriage, DM, hernia rep and myomectomy are the only significant discrete variables affecting the dependent variable (Tumor), this was done through using the Chi-square tests, with p-values less than 0.05, as shown in table 8.

The odds ratio shown in table 8, is computed as, (odds in exposed/odds on unexposed). For instance, for the previous marriage, the value 0.20 is calculated as ((320/321)/(30/6)) and can be interpreted as an estimate of the ratio of the odds, in the population, of previously married developing Malignant to the odds of a non-previous married this type of tumor. This odds ratio can be said that a previously married has 0.20 times the risk of a non-previous married of developing Malignant tumor. That means it has a good (negative) effect.

Also, the value 1.78 can be interpreted as an estimate of the ratio of the odds, in the population, of a diabetes developing Malignant to the odds of a non-diabetes this type of tumor. This odds ratio can be said that a diabetes has 1.78 times the risk of a non-diabetes of developing Malignant tumor. That means it has a bad (positive) effect.

Similar interpretation can be given for the hernia rep with positive effect and myomectomy with negative effect on developing Malignant tumor.

Table 7 t-tests for the continuous variables

| Tumor type | N  | Mean | SD    | t     | P-value |
|------------|----|------|-------|-------|---------|
| AGE        |    |      |       |       |         |
| Benign     | 327| 49.58| 10.558| -2.722| 0.007   |
| Malignant  | 350| 52.06| 12.923|       |         |
| Parity     |    |      |       |       |         |
| Benign     | 315| 4.81 | 3.224 | 2.974 | 0.003   |
| Malignant  | 332| 4.04 | 3.357 |       |         |
| Miscarriage|    |      |       |       |         |
| Benign     | 314| 0.69 | 1.171 | 3.874 | 0.000   |
| Malignant  | 332| 0.36 | 0.987 |       |         |

Table 8 χ²-tests for the categorical variables

| Tumor type | Malignant | Benign | Total | χ²     | P-value | Odds Ratio |
|------------|-----------|--------|-------|--------|---------|------------|
|            | n | % | n | % | n | % |       |        |        |
| Previous marriage | Yes | 320 | 49.90% | 321 | 50.10% | 641 | 100% | 15.238 | 0.000 | 0.20 |
|                 | No  | 30  | 83.30% | 6   | 16.70% | 36  | 100% |        |       |     |
| DM           | Yes | 95  | 62.50% | 57  | 37.50% | 152 | 100% | 9.533  | 0.002 | 1.78 |
|               | No  | 256 | 48.30% | 274 | 51.70% | 530 | 100% |        |       |     |
| hernia rep   | Yes | 9   | 90.00% | 1   | 10.00% | 10  | 100% | 6.033  | 0.014 | 8.68 |
|               | No  | 342 | 50.90% | 330 | 49.10% | 672 | 100% |        |       |     |
| myomectomy   | Yes | 4   | 25.00% | 12  | 75.00% | 16  | 100% | 4.595  | 0.032 | 0.31 |
|               | No  | 347 | 52.10% | 319 | 47.90% | 666 | 100% |        |       |     |

Predictive analytics

Predictive analytics mainly focus on developing a predictive model, that can be used to “predict” the Target, or the (Dependent) variable. We consider the Tumor type (i.e Benign/Malignant) is the Dependent variable. Since it is nominal, we must use the Logistic regression (Table 4). Stepwise logistic regression is used to include only the significant variables affecting the patient’s type and in descending order of importance. Table 9 summarizes the significance parameter estimates (B) and the odds ratios (Exp (B)), for the significant variables.

Citation: Abdelfattah Ezz H. Reclassifying inferential statistics into diagnostic and predictive statistics with an application on gynecologic cancer. Biom Biostat Int J. 2020;9(4):146–150. DOI: 10.15406/bbij.2020.09.00312
It is clear from the previous table that the Miscarriage is the 1st factor affecting the type of the tumor. It is also clear that, from the negative values for the coefficients of Miscarriage Previous Marriage, parity and myomectomy, and consequently the corresponding values of the odds ratios being less than 1, which means they decrease the chance of having a Malignant tumor. While the positive values for the coefficients of DM, hernia rep, and Age, and consequently the corresponding values of the odds ratios being more than 1, which means they increase the chance of having a Malignant tumor.

For example for the Age, the value of 1.019 of odds ratio (Exp(B)) means that with the increase of one year in age the risk of Malignant tumor is increased 1.019 times provided all other factors are kept constant. Since one year increase does not give any significant change, therefore, we can see the significant change after 10 years. This is calculated as:

\[ e^{1.019} = e^{0.019 	imes 10} \approx 1.21. \]

This indicates that with an increase of 10 years in age the risk of Malignant tumor increases 1.21 times.

**Perspective analytics**

Perspective analytics is a result of all the previous analytics and gives the –in advance- guides that can be used to “avoid” the problem. Here, the problem is to have a Malignant tumor. Based on the results obtained, we may support that Marriage, parity and myomectomy, will help decreasing the chance of having a Malignant tumor. While patients should avoid DM and hernia rep, that increase the chance of having a Malignant tumor.

**Acknowledgement**

None.

**Conflicts of interest**

None.

**References**

1. Abdelfattah Ezz H. A Modern way to Teach Statistics, with an application on a Medical Example. 4th Meeting, SASS, King Abdulaziz University – Jeddah. 2019.
2. Dipti NP, Atkotiya. KH. Cervical cancer prediction using data mining. International journal for research in applied science & engineering technology. 5.
3. Sunny Sharma. Cervical cancer stage prediction using decision tree approach of machine learning. International journal of advanced research in computer and communication engineering. 2016.
4. Eva C, Perez C, Cabrera S, et al. Molecular markers of endometrial carcinoma detected in uterine aspirates. Int J Cancer. 2011;129(10):2435-2444.
5. https://www.cdc.gov/cancer/gynecologic/basic_info/index.htm
6. David M, Technikum Wien FH. Support vector machines. The Interface to libsvm in package: e1071(2015):28.
7. Collins Y, Holcomb K, Champman-Davis E, et al. Gynecologic cancer disparities: a report from the Health Disparities Taskforce of the Society of Gynecologic Oncology. Gynecol Oncol. 2014;133(2):353-361.
8. Ramachandran P, Girija N, Bhuavaneswari T. Early detection and prevention of cancer using data mining techniques. International Journal of Computer Applications. 2014; 97(13).
9. Hanif M, Ahmad H, Abdelfattah Ezz H. Biostatistics For Health Students with Manual on Software Application. 2nd Edition, ISOSS, Lahore, Pakistan. 2014.