Statistics in Epidemiology

Epidemiology

Epidemiology is the study of the occurrence of disease or other health-related characteristics in human and in animal populations. Epidemiologists study the frequency of disease and whether the frequency differs across groups of people; such as, the cause-effect relationship between exposure and illness. Diseases do not occur at random; they have causes. Many diseases could be prevented if the causes were known. The methods of epidemiology have been crucial to identifying many causative factors which, in turn, have led to health policies designed to prevent disease, injury and premature death.

Basic Terms

Incidence is the number of new cases in a defined population within a specified period of time.[1]

Prevalence is the proportion of population who have specific characteristic in given time. It is disease occurrence or other factor related to health, the total number of individuals who have the condition at a particular time divided by the population at risk of having the condition at that time or midway through the period.

- Example: 400 people are tested for the common flue. 100 of them in the sample group are found to have the flue. Divide the 100 flue infected people by the total sample size, which is 400, the answer is the prevalence.

100/400= 1/4 or 1 out of 4 people or 25%.

Mortality determines how many people die in a certain time period. It can be measured with calculating the death rate: (Number of deaths during a specified period in the sample group)/(The total number of people in the sample group)

- Example: In the flue case 10 people died from the flue and 10 from a different source. So, divide the number of people who died by the total sample size: 20/400=1/20 or 0.5%, that is the mortality rate, notice that it is the TOTAL number of deaths, not just from the disease. For a more accurate measure use lethality rate.

Lethality of diseases is a ratio which is determined by the number of people who died in a certain time divided through the number of people who fell ill in the same time period. It is a description of how a disease can cause death.

- Example: Let’s take the flue case as an example. 10 people have died from the flue out of the 100 that were sick. So: 10/100=1/10 or in other works there is a 10% chance to die from this disease, that’s the lethality.

Diagnostic Tests

Diagnostic tests are performed in the aim of determining the presence of a certain disease or illness in a patient. The test may be carried out through performing procedures, such as various scans or merely on the basis of symptoms. Some examples of diagnostic tests include X-rays, biopsies, pregnancy tests, blood tests, results from physical examinations, etc.[1]

The results obtained from the test could be from either one f the 2 distinct main categories- positive or negative, where a positive result indicates the presence of the diseases. A positive or negative result can be subdivided further into true positives and negatives, and false positives and negative results. A true positive result is one that accurately determines the presence of the illness. On the contrary a false positive result indicates the presence of the disease in the patient; however, the disease is actually not present in the patient. A similar pattern is seen in true negative and false negative results. [1]

- True *positive*: the patient has the disease and the test is positive.
- *False positive*: the patient does not have the disease but the test is positive.
- *True negative*: the patient does not have the disease and the test is negative
- *False negative*: the patient has the disease but the test is negative.

Sensitivity and Specificity of Diagnostic Test Calculated from Fourfold table

The fourfold table is a type of contingency table which is a tabular cross-classification of data in which subcategories of one characteristic are indicated horizontally (in rows) and subcategories of another characteristic are indicated vertically (in columns) to test the characteristics between the two (the rows and the columns).

Diagnostic test result	True status Diseased	True status Non - diseased	Total
Positive	a	b	a+b
Negative	c	d	c+d
Total	a+c	b+d	a+b+c+d

- **Sensitivity** (also called the true positive rate) measures the proportion of all positives thatare correctly identified as positives.
- **Specificity** (also called the true negative rate) measures the proportion of all negatives that are correctly identified as negatives

In the fourfold table, the letters a, b, c, d symbolize the numbers in the fourfold table.

- a — stand for diseased individuals detected by the test.
- b — stand for healthy individuals detected by the test.
- c — stand for diseased individuals not detectable by the test.
- d — stand for healthy individuals negative by the test.

The basic statistics can be calculated form the fourfold table as follows:

The formula for sensitivity — a / a+c.

The formula for specificity — d / b+d.

predictive value of a positive test result — a / a+b.

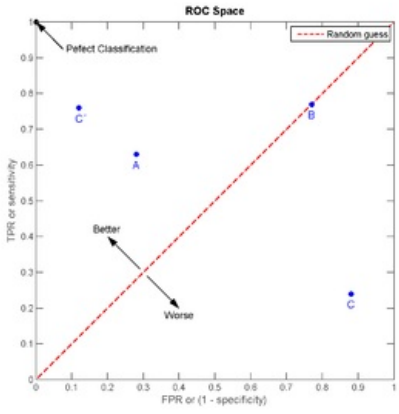
Predictive value of a negative test result — d / c+d.

Receiver Operating Characteristic (ROC) Curve

- Receiver Operating Characteristic (ROC curve) is a graphical plot that illustrates the performance of a binary classifier system (e.g. diagnosis test)

- The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

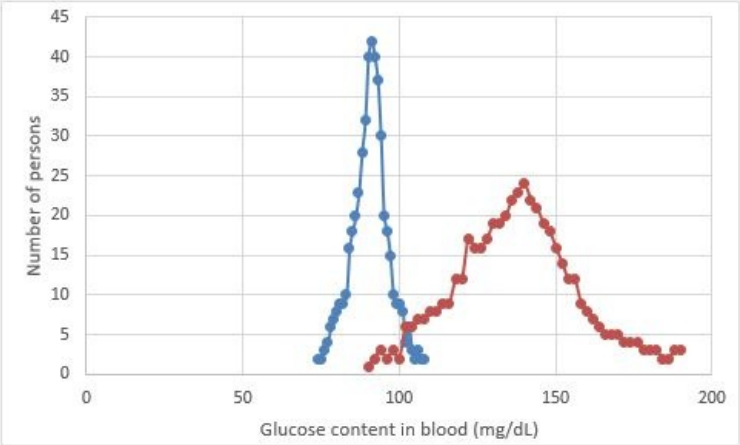
- ROC has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research.



ROC space

ROC curve example

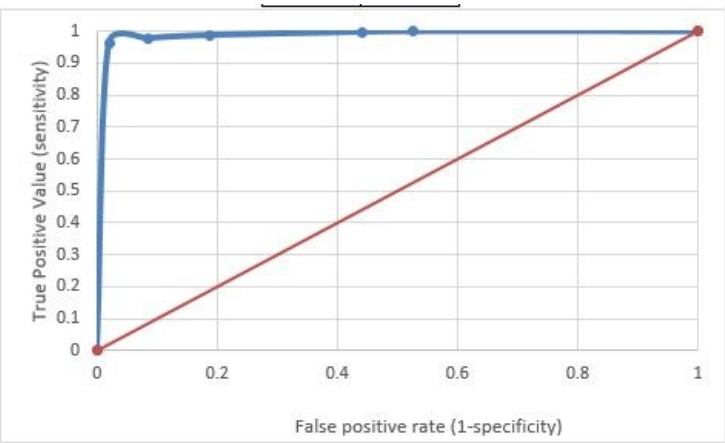
An *example of diabetes disease prediction using the blood glucose level was used to illustrated the way these curves are generated. In this example, it is supposed that the normal levels for a healthy person of glucose in blood are 70-110 mg/dL (average value 90 mg/dl) corresponding to condition negative, and the glucose concentration in blood of a non-healthy person is 90 to 180 mg/dL (corresponding to condition positive).* For this specific example, it is assumed that the sample of total population is 1000 members with 500 being non-healthy (condition positive) and 500 healthy (condition negative). The distribution of the healthy and non-healthy population is shown in the following graph.



From the graph, it is seen that the data are overlapping and therefore we are not able to distinguish between the two categories (diabetes and non-diabetes) with a 100% certainty. It is not clear from the graph what is the best threshold value for the categorization into the two categories (is it 95 mg/dL, 100 mg/dL or 105 mg/dL?). Different values of threshold will give a different sensitivity and specificity of the test. The ROC curve can help us to decide which threshold value would be the best for the particular situation.

Therefore, the parameters TP, FN, FP and TN, as well as the respective rates (TPR, FNR, FPR and TNR) are determined for five different cut-off (threshold) levels: 90, 92, 95, 100 and 105 mg/dL. The results of TPR and FPR for these exercises are presented the following table.

Threshold	FPR	TPR
90	0.980	0.962
92	0.916	0.978
95	0.812	0.988
100	0.558	0.998
105	0.474	1



As seen in The table, when sensitivity increases specificity decreases. Therefore the optimal threshold value should be set according to the situation. Some situations require a high sensitive tests (screening) and some high specific tests (In cases where the patient can be harmed with the upcoming treatment)

Screening and Confirmatory Test in Medicine

In the real world, you never have a test that is fully Sensitive and fully Specific. We are usually faced with a decision to use a test with high Sensitivity (and lower spec) or high Specificity (and lower Sensitivity). Usually a test with high sensitivity is used as the Initial Screening Test. Those that receive a positive result on the first test will be given a second test with high specificity that is used as the Confirmatory Test. In these situations, you need both tests to be positive to get a definitive diagnosis. Getting a single positive reading is not enough for a diagnosis as the individual tests have either a high chance of FP or a high chance of FN. For example, HIV is diagnosed using 2 tests. First an ELISA screening test is used and then a confirmatory Western Blot is used if the first test is positive. [1], [2]

There are also specific situations where having a high specificity or sensitivity is really important. Consider that you are trying to screen donations to a blood bank for blood borne pathogens. In this situation, you want a super high sensitivity, because you may infect anybody easily so the drawbacks of a false negative (spreading disease to a recipient) are way higher than the drawbacks of a false positive (throwing away 1 blood donation). Now consider you are testing a patient for the presence of a disease. This particular disease is treatable, but the treatment has very serious side effects (e.g. cancer treated by chemotherapy) . In this case, you want a test that has high specificity, because there are major drawbacks to a false positive. [1], [2]

Links

References

1. Porta, M. A Dictionary of Epidemiology. Oxford University Press
2. WikiLectures. Fourfold and Contingency Tables. 2014
3. Setiabudhi Times. The 10 Sources of Psychological Myths: Your Mythbusting Kit.2012
4. Lalkhen, A. McCluskey, A. Clinical tests: sensitivity and specificity.
5. PennState Eberly College of Science. Epidemiological Research Methods, 10.3 - Sensitivity, Specificity, Positive Predictive Value, and Negative Predictive Value
6. 4.6: Receiver Operating Characteristic Curves (author unknown) link: http://ebp.uga.edu/courses/Chapter%204%20-%20Diagnosis%20I/8%20-%20ROC%20curves.html
7. Department of Statistics, University of California, One Shields Ave, Davis, CA 95616, USA. Estimation of diagnostic-test sensitivity and specificity through Bayesian modelling. 2005

1. Porta Miquel, A dictionary of epidemiology, Oxford, sixth edition 2014.
2. Farlex Partner Medical Dictionary - Epidemiology, ROC analysis © Farlex 2012