Sensitivity and specificity

From Wikipedia, the free encyclopedia

Sensitivity and **specificity** are statistical measures of the performance of a binary classification test, also known in statistics as classification function. **Sensitivity** (also called the *true positive rate*, or the **recall rate** in some fields) measures the proportion of actual positives which are correctly identified as such (e.g. the percentage of sick people who are correctly identified as having the condition). **Specificity** (sometimes called the *true negative rate*) measures the proportion of negatives which are correctly identified as such (e.g. the percentage of healthy people who are correctly identified as not having the condition). These two measures are closely related to the concepts of type I and type II errors. A perfect predictor would be described as 100% sensitive (i.e. predicting all people from the sick group as sick) and 100% specific (i.e. not predicting anyone from the healthy group as sick); however, theoretically any predictor will possess a minimum error bound known as the Bayes error rate.

For any test, there is usually a trade-off between the measures. For example: in an airport security setting in which one is testing for potential threats to safety, scanners may be set to trigger on low-risk items like belt buckles and keys (low specificity), in order to reduce the risk of missing objects that do pose a threat to the aircraft and those aboard (high sensitivity). This trade-off can be represented graphically as a receiver operating characteristic curve.

Contents

Definitions

Imagine a study evaluating a new test that screens people for a disease. Each person taking the test either has or does not have the disease. The test outcome can be positive (predicting that the person has the disease) or negative (predicting that the person does not have the disease). The test results for each subject may or may not match the subject's actual status. In that setting:

- True positive: Sick people correctly diagnosed as sick
- False positive: Healthy people incorrectly identified as sick

- True negative: Healthy people correctly identified as healthy
- False negative: Sick people incorrectly identified as healthy

In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected

Let us define an experiment from **P** positive instances and **N** negative instances for some condition. The four outcomes can be formulated in a 2×2 *contingency table* or *confusion matrix*, as follows:

Terminology and derivations from a confusion matrix

true positive (TP)
eqv. with hit
true negative (TN)
eqv. with correct rejection
false positive (FP)
eqv. with false alarm, Type I error
false negative (FN)
eqv. with miss, Type II error

---

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall
$$TPR = TP/P = TP/(TP+FN)$$
specificity (SPC) or True Negative Rate
$$SPC = TN/N = TN/(FP+TN)$$
precision or positive predictive value (PPV)
$$PPV = TP/(TP+FP)$$
negative predictive value (NPV)
$$NPV = TN/(TN+FN)$$
fall-out or false positive rate (FPR)
$$FPR = FP/N = FP/(FP+TN) = 1 - SPC$$
false discovery rate (FDR)
$$FDR = FP/(TP+FP) = 1 - PPV$$

---

accuracy (ACC)
$$ACC = (TP+TN)/(P+N)$$
F1 score
is the harmonic mean of precision and sensitivity
$$F1 = 2TP/(2TP+FP+FN)$$
Matthews correlation coefficient (MCC)
$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Source: Fawcett (2006).

<table>
<tr><td colspan="2"></td><td colspan="2">Condition
(as determined by "Gold standard")</td><td></td></tr>
<tr><td colspan="2"></td><td>Condition positive</td><td>Condition negative</td><td></td></tr>
<tr><td rowspan="2">Test outcome</td><td>Test outcome positive</td><td>True positive</td><td>False positive
(Type I error)</td><td>Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$</td></tr>
<tr><td>Test outcome negative</td><td>False negative
(Type II error)</td><td>True negative</td><td>Negative predictive value = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$</td></tr>
<tr><td colspan="2"></td><td>Sensitivity = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>Specificity = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Accuracy</td></tr>
</table>

Sensitivity

Sensitivity relates to the test's ability to identify a condition correctly. Consider the example of a medical test used to identify a disease. Sensitivity of the test is the proportion of people known to have the disease, who test positive for it. Mathematically, this can be expressed as:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$= \frac{\text{number of true positives}}{\text{total number of sick individuals in population}}$$

$$= \text{probability of a positive test, given that the patient is ill}$$

Negative result in a test with high sensitivity is useful for ruling out disease. A high sensitivity test is reliable when its result is negative, since it rarely misdiagnoses those who have the disease. A test with 100% sensitivity will recognize all patients with the disease by testing positive. A negative test result would definitively *rule out* presence of the disease in a patient.

Positive result in a test with high sensitivity is not useful for ruling in disease. Suppose a 'bogus' test kit is designed to show only one reading, positive. When used on diseased patients, all patients test positive, giving the test 100% sensitivity. However, sensitivity by definition does not take into account false positives. The bogus test also returns positive on all healthy patients, giving it a false positive rate of 100%, rendering it useless for diagnosing or "ruling in" the disease.

Sensitivity is not the same as the precision or positive predictive value (ratio of true positives to combined true and false positives), which is as much a statement about the proportion of actual positives in the population being tested as it is about the test.

The calculation of sensitivity does not take into account indeterminate test results. If a test cannot be repeated, indeterminate samples either should be excluded from the analysis (the number of exclusions should be stated when quoting sensitivity) or can be treated as false negatives (which gives the worst-case value for sensitivity and may therefore underestimate it).

A test with high sensitivity has a low type II error rate. In non-medical contexts, sensitivity is sometimes called recall.

Specificity

Specificity relates to the test's ability to exclude a condition correctly. Consider the example of a medical test for diagnosing a disease. Specificity of a test is the proportion of healthy patients known not to have the disease, who will test negative for it. Mathematically, this can also be written as:

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

$$= \frac{\text{number of true negatives}}{\text{total number of well individuals in population}}$$

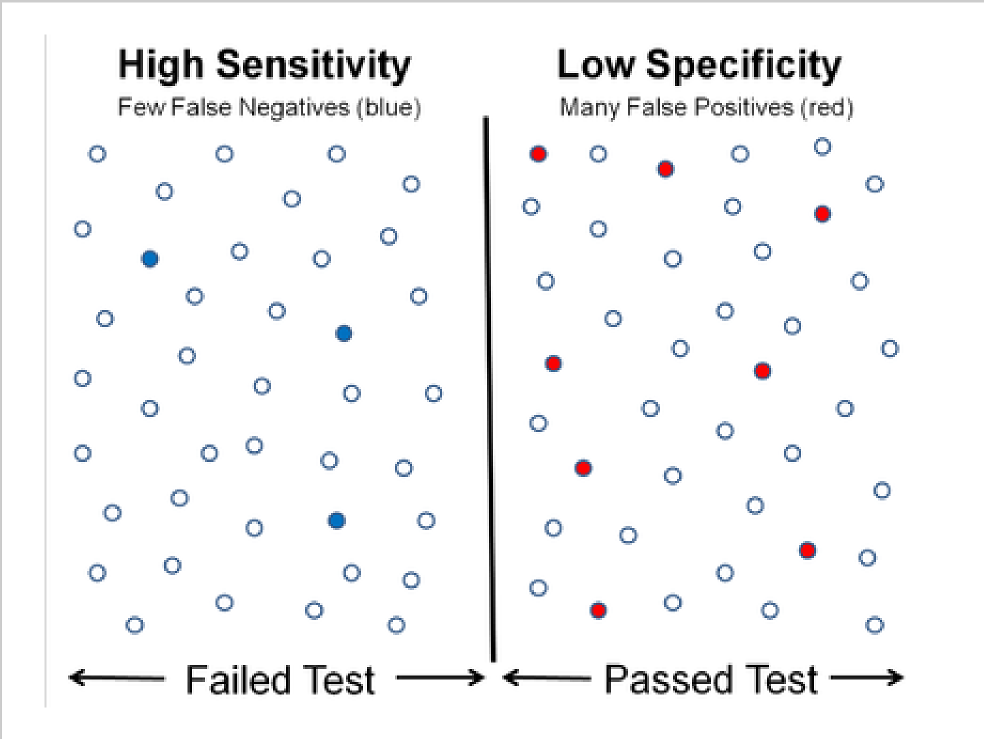
$$= \text{probability of a negative test given that the patient is well}$$

Positive result in a test with high specificity is useful for ruling in disease. The test rarely gives positive results in healthy patients. A test with 100% specificity will read negative, and accurately exclude disease from all healthy patients. A positive result will highlight a high probability of the presence of disease.[1]

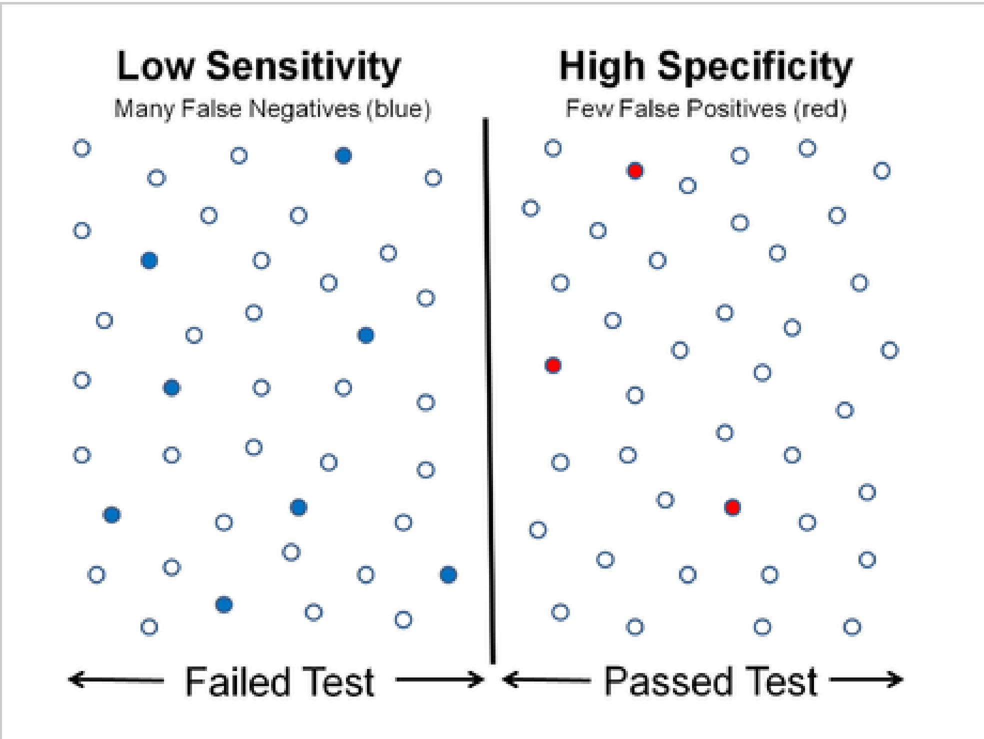
Negative result in a test with high specificity is not useful for ruling out disease. Assume a 'bogus' test is designed to read only negative. This is administered to healthy patients, and reads negative on all of them. This will give the test a specificity of 100%. Specificity by definition does not take into account false negatives. The same test will also read negative on diseased patients, therefore it has a false negative rate of 100%, and will be useless for ruling out disease.

A test with a high specificity has a low type I error rate.

Graphical illustration



High sensitivity and low specificity



Low sensitivity and high specificity

Medical examples

In medical diagnostics, test sensitivity is the ability of a test to correctly identify those with the disease (true positive rate), whereas test specificity is the ability of the test to correctly identify those without the disease (true negative rate). If 100 patients known to have a disease were tested, and 43 test positive, then the test has 43% sensitivity. If 100 with no disease are tested and 96 return a negative result, then the test has 96% specificity. Sensitivity and specificity are prevalence-independent test characteristics, as their values are intrinsic to the test and do not depend on the disease prevalence in the population of interest.[2] Positive and negative predictive values, but not sensitivity or specificity, are values influenced by the prevalence of disease in the population that is being tested.

Misconceptions

It is often claimed that a highly specific test is effective at ruling in a disease when positive, while a highly sensitive test is deemed effective at ruling out a disease when negative.[3][4] This has led to the widely used mnemonics SPIN and SNOUT, according to which a highly SPecific test, when Positive, rules IN disease (SP-P-IN), and a highly 'SeNsitive' test, when Negative rules OUT disease (SN-N-OUT). Both rules of thumb are, however, inferentially misleading, as the diagnostic power of any test is determined by *both* the sensitivity and specificity.[5][6][7]

Sensitivity_index

The **sensitivity index** or ***d'*** (pronounced 'dee-prime') is a statistic used in signal detection theory. It provides the separation between the means of the signal and the noise distributions, compared against the standard deviation of the noise distribution. For normally distributed signal and noise with mean and standard deviations μ_S and σ_S, and μ_N and σ_N, respectively, d' is defined as:

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}(\sigma_S^2 + \sigma_N^2)}}$$ [8]

An estimate of d' can be also found from measurements of the hit rate and false-alarm rate. It is calculated as:

d' = *Z*(hit rate) - *Z*(false alarm rate),[9]

where function $Z(p)$, $p \in [0,1]$, is the inverse of the cumulative Gaussian distribution.

d' is a dimensionless statistic. A higher *d'* indicates that the signal can be more readily detected.

Worked example

A worked example

A diagnostic test with sensitivity 67% and specificity 91% is applied to 2030 people to look for a disorder with a population prevalence of 1.48%

		Patients with bowel cancer (as confirmed on endoscopy)		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Related calculations

- False positive rate (α) = type I error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%
- False negative rate (β) = type II error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) = 33%
- Power = sensitivity = 1 − β
- Likelihood ratio positive = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4
- Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

Hence with large numbers of false positives and few false negatives, a positive screen test is in itself poor at confirming the disorder (PPV = 10%) and further investigations must be undertaken; it did, however, correctly identify 66.7% of all cases (the sensitivity). However as a screening test, a negative result is very good at reassuring that a patient does not have the disorder (NPV = 99.5%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

Estimation of errors in quoted sensitivity or specificity

Sensitivity and specificity values alone may be highly misleading. The 'worst-case' sensitivity or specificity must be calculated in order to avoid reliance on experiments with few results. For example, a particular test may easily show 100% sensitivity if tested against the gold standard four times, but a single additional test against the gold standard that gave a poor result would imply a sensitivity of only 80%. A common way to do this is to state the binomial proportion confidence interval, often calculated using a Wilson score interval.

Confidence intervals for sensitivity and specificity can be calculated, giving the range of values within which the correct value lies at a given confidence level (e.g. 95%).[10]

Terminology in information retrieval

In information retrieval, the positive predictive value is called **precision**, and sensitivity is called **recall**.

The F-score can be used as a single measure of performance of the test. The F-score is the harmonic mean of precision and recall:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

In the traditional language of statistical hypothesis testing, the sensitivity of a test is called the statistical power of the test, although the word *power* in that context has a more general usage that is not applicable in the present context. A sensitive test will have fewer Type II errors.

See also

- Accuracy and precision
- Brier score
- Confusion matrix
- Detection theory
- F-score
- Gain (information retrieval)
- Likelihood ratios
- Matthews correlation coefficient
- OpenEpi software program
- Precision and recall
- Receiver operating characteristic or ROC curve
- Selectivity
- Sensitivity index
- Statistical significance
- Youden's J statistic
- Uncertainty coefficient, aka Proficiency

References

1. ^ "SpPins and SnNouts" (http://www.cebm.net/index.aspx?o=1042). Centre for Evidence Based Medicine (CEBM). Retrieved 26 December 2013.
2. ^ Mangrulkar, Rajesh. "Diagnostic Reasoning I and II" (http://open.umich.edu/education/med/m1/patientspop-decisionmaking/2010/materials). Retrieved 24 January 2012.
3. ^ Michigan State University Evidence Based Medicine resource (http://omerad.msu.edu/ebm/Diagnosis/Diagnosis4.html)
4. ^ Emory University Medical School Evidence Based Medicine course (http://www.med.emory.edu/EMAC/curriculum/diagnosis/sensand.htm)
5. ^ Baron, JA (Apr–Jun 1994). "Too bad it isn't true.....". *Medical decision making : an international journal of the Society for Medical Decision Making* **14** (2): 107. doi:10.1177/0272989X9401400202 (http://dx.doi.org/10.1177%2F0272989X9401400202). PMID 8028462 (//www.ncbi.nlm.nih.gov/pubmed/8028462).
6. ^ Boyko, EJ (Apr–Jun 1994). "Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn?". *Medical decision making : an international journal of the Society for Medical Decision Making* **14** (2): 175–179. doi:10.1177/0272989X9401400210 (http://dx.doi.org/10.1177%2F0272989X9401400210). PMID 8028470 (//www.ncbi.nlm.nih.gov/pubmed/8028470).
7. ^ Pewsner, D; Battaglia, M; Minder, C; Marx, A; Bucher, HC; Egger, M (Jul 24, 2004). "Ruling a diagnosis in or out with "SpPIn" and "SnNOut": a note of caution" (//www.ncbi.nlm.nih.gov/pmc/articles/PMC487735). *BMJ (Clinical research ed.)* **329** (7459): 209–13. doi:10.1136/bmj.329.7459.209 (http://dx.doi.org/10.1136%2Fbmj.329.7459.209). PMC 487735 (//www.ncbi.nlm.nih.gov/pmc/articles/PMC487735). PMID 15271832

(//www.ncbi.nlm.nih.gov/pubmed/15271832).
8. ^ Gale, SD; Perkel, DJ (Jan 20, 2010). "A basal ganglia pathway drives selective auditory responses in songbird dopaminergic neurons via disinhibition" (//www.ncbi.nlm.nih.gov/pmc/articles/PMC2824341). *The Journal of neuroscience : the official journal of the Society for Neuroscience* **30** (3): 1027–1037. doi:10.1523/JNEUROSCI.3585-09.2010 (http://dx.doi.org/10.1523%2FJNEUROSCI.3585-09.2010). PMC 2824341 (//www.ncbi.nlm.nih.gov/pmc/articles/PMC2824341). PMID 20089911 (//www.ncbi.nlm.nih.gov/pubmed/20089911).
9. ^ Macmillan, Neil A.; Creelman, C. Douglas (15 September 2004). *Detection Theory: A User's Guide* (http://books.google.com/books?id=hDX65v9bReYC). Psychology Press. p. 7. ISBN 978-1-4106-1114-7.
10. ^ Online calculator of confidence intervals for predictive parameters (http://www.medcalc.org/calc/diagnostic_test.php)

Further reading

- Altman DG, Bland JM (1994). "Diagnostic tests. 1: Sensitivity and specificity" (http://www.bmj.com/cgi/content/full/308/6943/1552). *BMJ* **308** (6943): 1552. doi:10.1136/bmj.308.6943.1552 (http://dx.doi.org/10.1136%2Fbmj.308.6943.1552). PMC 2540489 (//www.ncbi.nlm.nih.gov/pmc/articles/PMC2540489). PMID 8019315 (//www.ncbi.nlm.nih.gov/pubmed/8019315).
- Loong T (2003). "Understanding sensitivity and specificity with the right side of the brain" (//www.ncbi.nlm.nih.gov/pmc/articles/PMC200804). *BMJ* **327** (7417): 716–719. doi:10.1136/bmj.327.7417.716 (http://dx.doi.org/10.1136%2Fbmj.327.7417.716). PMC 200804 (//www.ncbi.nlm.nih.gov/pmc/articles/PMC200804). PMID 14512479 (//www.ncbi.nlm.nih.gov/pubmed/14512479).

External links

- Vassar College's Sensitivity/Specificity Calculator (http://faculty.vassar.edu/lowry/clin1.html)