

## Review

Data for assessing the sensitivity and specificity of a test are usually of the form

test result	disease category	
	diseased (+)	nondiseased (-)
+	$A$	$B$
-	$C$	$D$

**Sensitivity:** is the proportion of diseased people who would be correctly classified (estimated by  $Sens = A/(A + C)$ ).

**Specificity:** is the proportion of nondiseased people who would be correctly classified (estimated by  $Spec = D/(B + D)$ ).

**The prevalence:** of a disease is the percent of the population with the disease (estimated by  $R = (A + C)/(A + B + C + D)$ ). Note that a random sample is required to estimate prevalence.

**Positive Predictive Value:** is the proportion of people who tested positive that truly are positive. (estimated by  $PPV = A/(A + B)$ ).

**Negative Predictive Value:** is the proportion of people who tested negative that truly are negative. (estimated by  $NPV = D/(C + D)$ ).

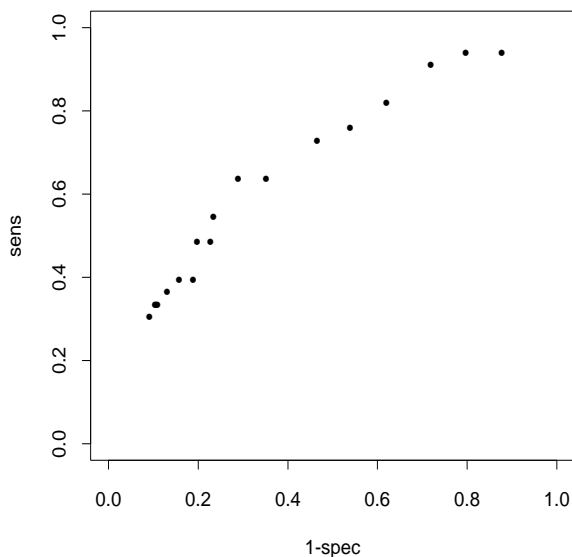
**False Negative:** The probability of a false negative is the probability of testing negative given a truly positive condition.

**False Positive:** The probability of a false positive is the probability of testing positive given a truly negative condition.

Serum Creat. (mg %)	Sens	Spec
1.2	0.939	0.123
1.3	0.939	0.203
1.4	0.909	0.281
1.5	0.818	0.380
1.6	0.758	0.461
1.7	0.727	0.535
1.8	0.636	0.649
1.9	0.636	0.711
2.0	0.545	0.766
2.1	0.485	0.773
2.2	0.485	0.803
2.3	0.394	0.811
2.4	0.394	0.843
2.5	0.364	0.870
2.6	0.333	0.891
2.7	0.333	0.894
2.8	0.333	0.896
2.9	0.303	0.909

Sensitivity and Specificity of serum creatinine level for predicting transplant rejection.

An ROC curve is a graph that plots the sensitivity as a function of the false positive (1 - spec) probability.



## More Review

A **random experiment** is an experiment for which the outcome cannot be predicted with certainty, but all possible outcomes can be identified prior to its performance, and it may be repeated under the same conditions.

The set of all possible outcomes of a random experiment is called the **sample space**, denoted  $\Omega$ ; subsets of the sample space are events, often denoted by  $A$  or  $A_k$  for  $k = 1, 2, \dots, n$ .

**Things to know about Probability:** Let  $\Omega$  be the sample space for a probability measure  $P$ .

- $0 \leq P(A) \leq 1$ , for all events  $A$ , and  $P(\Omega) = 1$ .
- $P(\emptyset) = 0$
- For each  $A \subset \Omega$ ,  $P(A) = 1 - P(\bar{A}) = 1 - P(A^c)$ .
- If  $A_1 \subset A_2 \subset \Omega$ ,  $P(A_1) \leq P(A_2)$
- Two events  $A_1$  and  $A_2$  are mutually exclusive if they cannot both happen at the same time :  $(A_1 \cap A_2 = \emptyset)$ . In this case  $P(A_1 \cap A_2) = 0$ .
- Two events are independent if  $P(A_1 \cap A_2) = P(A_1)P(A_2)$

### Things to know about Probability (continued):

7. For any events  $A_1$  and  $A_2$ ,

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

(note for mutually exclusive events, the probability of the union is the sum of the probabilities - this extends to more than just two groups - see notes from last time. This is sometimes called the **addition law of probability**).

8. For any events  $A_1$  and  $A_2$ ,

$$P(A_2|A_1) = \frac{P(A_2 \cap A_1)}{P(A_1)}$$

(note that for mutually exclusive events,  $P(A_2|A_1) = 0$ ; for independent events  $P(A_2|A_1) = P(A_2)$ ).

9. The above can be written as what is sometimes called the **multiplication rule of probability**:

$$P(A_2 \cap A_1) = P(A_2|A_1)P(A_1).$$

10. **Law of Total Probability**: Suppose events  $A_1, A_2, \dots, A_n$  are mutually exclusive and also exhaustive - exhaustive means  $\cup_{i=1}^n A_i = \Omega$ . Then,

$$P(B) = \sum_{i=1}^n P(A_i \cap B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

$$\begin{aligned} P(D^+|test^+) &= \frac{P(D^+ \cap test^+)}{P(test^+)} \\ &= \frac{P(test^+|D^+)P(D^+)}{P((test^+ \cap D^+) \cup (test^+ \cap D^-))} \\ &= \frac{P(test^+|D^+)P(D^+)}{P(test^+ \cap D^+) + P(test^+ \cap D^-)} \\ &= \frac{P(test^+|D^+)P(D^+)}{P(test^+|D^+)P(D^+) + P(test^+|D^-)P(D^-)} \\ &= \frac{Sens \cdot x}{Sens \cdot x + (1 - Spec) \cdot (1 - x)} \end{aligned}$$

where  $x = P(D^+)$  is the disease prevalence.

Note that a person randomly selected from a population (according to the probabilities estimated in our hypothetical study) has an 11% chance of being HIV positive. This probability (0.11) is called the *prior* probability.

In contrast, the *posterior* probability takes into account additional information. In this case, the additional information is the result of an HIV test.  $P(HIV^+|test^+)$  is a *posterior* probability.

Relationships between Sens, Spec, PPV, NPV, FN and FP.

Since 9.1 % of the truly positive people tested negative (1,000 out of 11,000), then  $(100 - 9.1 = 90.9)$  % of these truly positive people tested positive. Note that  $10,000/11,000 = 0.909$ .

Prob(false neg) = 0.091 and Sens is  $1 - 0.091 = 0.909$ .

Convince yourself that Spec =  $1 - PFP$  where PFP is the probability of a false positive.

### Bayes Rule

The example above is an application of Bayes Rule.

Bayes Rule: If  $A_1, A_2, \dots, A_n$  are  $n$  mutually exclusive events ( $A_i \cap A_j = \emptyset$  for  $i \neq j$ ) that are also exhaustive ( $\cup_{i=1}^n A_i = \Omega$ ), then for any event  $B$  where  $P(B) > 0$ ,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}$$

Suppose that a 60 year old man who has never smoked cigarettes presents to a physician with symptoms consisting of a chronic cough and occasional breathlessness. The physician expects that the patient is either if normal ( $A_1$ ), has lung cancer ( $A_2$ ), or has sarcoidosis ( $A_3$ ) and a biopsy is ordered. Suppose results of biopsy are positive or negative and that a positive result is consistent with either lung cancer or sarcoidosis. Outcomes are:

$$(A_1 \cap test^+), (A_1 \cap test^-), (A_2 \cap test^+), \dots, (A_3 \cap test^-)$$

What is  $P(A_1|test^+)$ ?

Note that  $(A_i \cap test^+) \cup (A_i \cap test^-) = A_i$  for  $i = 1, 2, 3$  are mutually exclusive events and that  $\cup_{i=1}^3 A_i = \Omega$ .

Bayes rule!

$$P(A_1|test^+) = \frac{P(test^+|A_1)P(A_1)}{\sum_{i=1}^3 P(test^+|A_i)P(A_i)}$$

How would  $P(A_1|test^+)$  change if the man was a smoker? In this case,  $P(A_1) = 0.98$ ,  $P(A_2) = 0.015$ ,  $P(A_3) = 0.005$ .

Convince yourself that if the man was a smoker,  $P(A_1|test^+) = 0.052$ ,  $P(A_2|test^+) = 0.711$ , and  $P(A_3|test^+) = 0.237$ .

Now, lung cancer is the most likely diagnosis.

Suppose that  $P(test^+|A_1) = 0.001$ ,  $P(test^+|A_2) = 0.9$  and  $P(test^+|A_3) = 0.9$  and that in 60 year old non-smoking men,  $P(A_1) = 0.99$ ,  $P(A_2) = 0.001$  and  $P(A_3) = 0.009$ .

The first set of probabilities could be obtained from clinical experience. The second set would have to be obtained from age-sex-smoking specific prevalence rates for the conditions in question.

$$\begin{aligned} P(A_1|test^+) &= \frac{P(test^+|A_1)P(A_1)}{\sum_{i=1}^3 P(test^+|A_i)P(A_i)} \\ &= \frac{0.001 \cdot 0.99}{0.001 \cdot 0.99 + 0.9 \cdot 0.001 + 0.9 \cdot 0.009} \\ &= 0.099 \end{aligned}$$

We could work out  $P(A_2|test^+) = 0.090$  and  $P(A_3|test^+) = 0.811$

Although the unconditional (prior) probability of sarcoidosis is very low (0.009), the conditional probability of the disease given these symptoms and this age/sex/smoking group is 0.811.

Consider a manuscript published in the *Journal of Ultrasound in Medicine* (1995).

Q1: Using the information in Table 3 of that paper, what is the probability that a 42 year old will have a child with autosomal trisomy? How does that probability change if you know the woman has had a test indicating that trisomy is not present?

Q2: What is the probability that a 35 year old woman will have a child with autosomal trisomy?

In the manuscript, sonographic scores were reviewed for 97 trisomic ( $T^+$ ) and 694 non-trisomic ( $T^-$ ) infants. These infants were not randomly sampled from the population. A positive test implies trisomy was detected. Of the 97, 83 tested positive; of the 694, 606 tested negative.

Sonographic test	trisomy		total
	( $T^+$ )	( $T^-$ )	
+	83	88	171
-	14	606	620
total	97	694	791

$$Sens = \frac{83}{97} = 0.85567$$

$$Spec = \frac{606}{694} = 0.87319$$

Q: What is the probability of a 42 year old woman having a child with trisomy given she has had a normal sonogram ?

$$P(T^+_{42}|test^-) = \frac{P(test^-|T^+_{42})P(T^+_{42})}{P(test^-|T^+_{42})P(T^+_{42}) + P(test^-|T^-_{42})P(T^-_{42})}$$

$P(test^-|T^+_{42})$  and  $P(test^-|T^-_{42})$  are taken from the study data given in the table shown previously (where sensitivity and specificity are calculated). One could argue that these probabilities are NOT affected by a woman's age, they only depend on the accuracy of the test.

$P(T^-_{42})$  and  $P(T^+_{42})$  are affected by a woman's age. Good estimates are available in the literature. These estimates are given in Table 3 of the manuscript.

$$\frac{(1 - 0.85567) \times 0.0306}{(1 - 0.85567) \times 0.0306 + (0.87319) \times 0.9694} = 0.00519$$

Q: What is the probability of a 42 year old woman having a child with trisomy,  $P(T^+_{42})$  ?

We cannot get an estimate from the table on the previous page for two main reasons: (1) The table incorporates infants from women of varying ages and (2) The table is NOT a random sample of 791 infants.

We can get an estimate of  $P(T^+_{42})$  from previous studies. Data from other studies is shown in Table 3 of the manuscript.

The study was done to evaluate the effectiveness of the sonogram to accurately detect trisomy. To do this, cases of known trisomy were obtained and how well the test performed was recorded. Similarly, cases of known non-trisomy were obtained and test performance was recorded. Using this information, sensitivity and specificity estimates were obtained. Because this was not a random sample of a population, an accurate estimate of the prevalence could not be obtained. This information was obtained from the literature. \*\*\*\* Bayes rule was used to combine the estimates of sensitivity and specificity with the estimate of prevalence to answer the question: "What is the probability of a 42 year old woman having a child with trisomy given she has had a normal sonogram ?".

We could answer the question for any child bearing aged woman (age  $n$ ) by adjusting our estimate of the prevalence.

$$P(T^+_n|test^-) = \frac{P(test^-|T^+_n)P(T^+_n)}{P(test^-|T^+_n)P(T^+_n) + P(test^-|T^-_n)P(T^-_n)}$$

We just looked at

$$P(d^+|test^-) = \frac{P(test^-|d^+)P(d^+)}{P(test^-|d^+)P(d^+) + P(test^-|d^-)P(d^-)}$$

Again, two reasons Bayes rule is important:

It allows you to incorporate extra information into probability estimates. For example, the *prior* probability  $P(d^+)$  is adjusted to account for additional information resulting in the *posterior* probability,  $P(d^+|test^+)$

It is a formula that relates conditional probabilities. Many times, certain conditional probabilities are well known (e.g. in the literature). You can use these to obtain conditional probabilities that you're interested in.

Other important conditional probabilities:

#### Relative Risk and Odds Ratio

The **relative risk** (RR) is the probability that a member of an exposed group will develop a disease relative to the probability that a member of an unexposed group will develop that same disease.

$$RR = \frac{P(disease|exposed)}{P(disease|unexposed)}$$

If an event takes place with probability  $p$ , the odds in favor of the event are  $\frac{p}{1-p}$  to 1.  $p = \frac{1}{2}$  implies 1 to 1 odds;  $p = \frac{2}{3}$  implies 2 to 1 odds.

In this class, the **odds ratio** (OR) is the odds of disease among exposed individuals divided by the odds of disease among unexposed.

$$OR = \frac{P(disease|exposed)/(1 - P(disease|exposed))}{P(disease|unexposed)/(1 - P(disease|unexposed))}$$