R doc: Introduction to probability computation and applications

Josep L. Carrasco
Bioestadística. Departament de Fonaments Clínics
Universitat de Barcelona

Probability computation

Case example

Load the file pumps.txt corresponding to the example of the infusion pumps.

```
pump=read.table("pumps.txt",sep="*",header=T)
```

Let's use only the data at $dosi = 2$

```
pump=pump[pump$dosi==2,]
```

Generate a contingency table showing the frequencies of the variables "Number of alarms" (nalarma) and "Pump origin" (origen).

```
tab<-table(pump$nalarma,pump$origen)
addmargins(tab)
```

```
        A   B   C   N   S Sum
  0     0   2   4  13   5  24
  1     4   4   6  12   8  34
  2     5   6   4   5   4  24
  3     5   7   4   1   1  18
  4     5   2   3   0   3  13
  5     2   1   1   0   0   4
  6     0   0   1   0   0   1
  Sum  21  22  23  31  21 118
```

Probabilities

Compute the following probabilities using the table. Round the results to 4 decimal places.

- Probability that a pump comes from origin A.

$P\left(O="A"\right)=\frac{21}{118}$

```r
round(21/118,digits=4)
```

```
[1] 0.178
```

- Probability that a pump sounds more than 3 times.

$P\left(N>3\right)=P\left(N\geq 4\right)=\frac{13+4+1}{118}=\frac{18}{118}$

```r
round(18/118,digits=4)
```

```
[1] 0.1525
```

- Probability that an alarm sounds more than 3 times AND its origin is A.

$P\left(N>3\cap O="A"\right)=\frac{5+2+0}{118}=\frac{7}{118}$

```r
round(7/118,digits=4)
```

```
[1] 0.0593
```

- Probability that an alarm sounds more than 3 times OR its origin is A.

$P\left(N>3\cup O="A"\right)=P\left(N>3\right)+P\left(O="A"\right)-P\left(N>3\cap O="A"\right)=\frac{18}{118}+\frac{21}{118}-\frac{7}{118}=\frac{32}{118}$

```r
round(32/118,digits=4)
```

```
[1] 0.2712
```

- Probability that an alarm sounds more than 3 times IF its origin is A.

$P\left(N>3|O="A"\right)=\frac{P(N>3\cap O="A")}{P(O="A")}=\frac{\frac{7}{118}}{\frac{21}{118}}=\frac{7}{21}$

```
round(7/21,digits=4)
```

```
[1] 0.3333
```

Assessment of diagnostic test ability

Case example

A sample of 100 subjects with fever were enrolled in an observational assay. The aim of the research was to quickly detect a potential infection. Gold standard procedure involves a culture study that may last for days. To shorten the time up to diagnosis, clinicians proposed one continuous and one binary marker (positive or negative).

Let's load the data.

```
dades=read.table("dades.txt",sep="\t",header=T)
```

The data set contains three columns:

- Outcome. It indicates if the subject was infected (I) or not (NI).
- Test1. Continuous marker.
- Test2. Binary marker (P=Positive; N=Negative).

Accuracy and utility indexes

We are going to use the *ThresholdROC* package to compute the accuracy and utility indexes.

```
library(ThresholdROC)
```

Let's use the binary test. Before computing the indexes, it is necessary to build a contingency table with the test and the outcome.

```
mytab=table(dades$Test2,dades$Outcome)
mytab
```

```
     I NI
  0 22 23
  1 50  5
```

The test are on the rows and the outcome on the columns of the table.

To estimate the indexes we will use the function *diagnostic*. To explore the arguments of this function run *?diagnostic*.

Main arguments are

- *tab*. A table that brings the frequencies of true and false positives / negatives. The configuration must be:

TP	FP
FN	TN

That means, positive results on the first row, outcome present on the first column. In our case, notice that the outcome is correctly distributed on the columns. However, the positive results of the test are on the second row. To change the order of the rows just run:

```
mytab=table(dades$Test2,dades$Outcome)[2:1,]
mytab
```

```
    I NI
  1 50  5
  0 22 23
```

- *casecontrol*. Were data collected in a case-control study? The default option is FALSE, what means the prevalence (proportion of outcome present) from the table can be used to estimate the predictive values. Otherwise, it should be indicated as TRUE and a value for the prevalence has to be given using the argument *p*.

Let's compute the accuracy and utility indexes.

```
diagnostic(mytab)
```

```
                  Estim. Low.lim(95%) Up.lim(95%)
Sensitivity    0.6944444    0.5732542   0.7946898
Specificity    0.8214286    0.6241744   0.9323251
Pos.Pred.Val.  0.9090909    0.7929072   0.9660460
Neg.Pred.Val.  0.5111111    0.3597755   0.6605712
LR+            3.8888889    1.7316367   8.7336202
LR-            0.3719807    0.2521812   0.5486914
Odds ratio    10.4545455    3.5171643  31.0754659
Youden index   0.5158730    0.3385444   0.6932016
Accuracy       0.7300000    0.6303929   0.8116352
Error rate     0.2700000    0.1883648   0.3696071
```

The estimated values are on the first column. Second and third columns correspond to the 95% confidence intervals (this concept has not been treated yet in the course).

Let's now to assess the continuous marker named as *Test1* in the data. The means of the marker in function of the Outcome are:

```
by(dades$Test1,dades$Outcome,mean)
```

```
dades$Outcome: I
[1] 6.56
------------------------------------------------------------
dades$Outcome: NI
[1] 4.051429
```

The mean of the infected subjects is greater than that of non-infected subjects. Therefore it makes sense to use larger values of Test1 as indicative of infection.

Researchers set a value of 5 as a threshold, so that a subject with a value greater of 5 will be labelled as "positive". Let's generate such a variable.

```
dades$Test3=dades$Test1>5
```

Let's assess the new binary marker.

```
mytab2=table(dades$Test3,dades$Outcome)[2:1,]
diagnostic(mytab2)
```

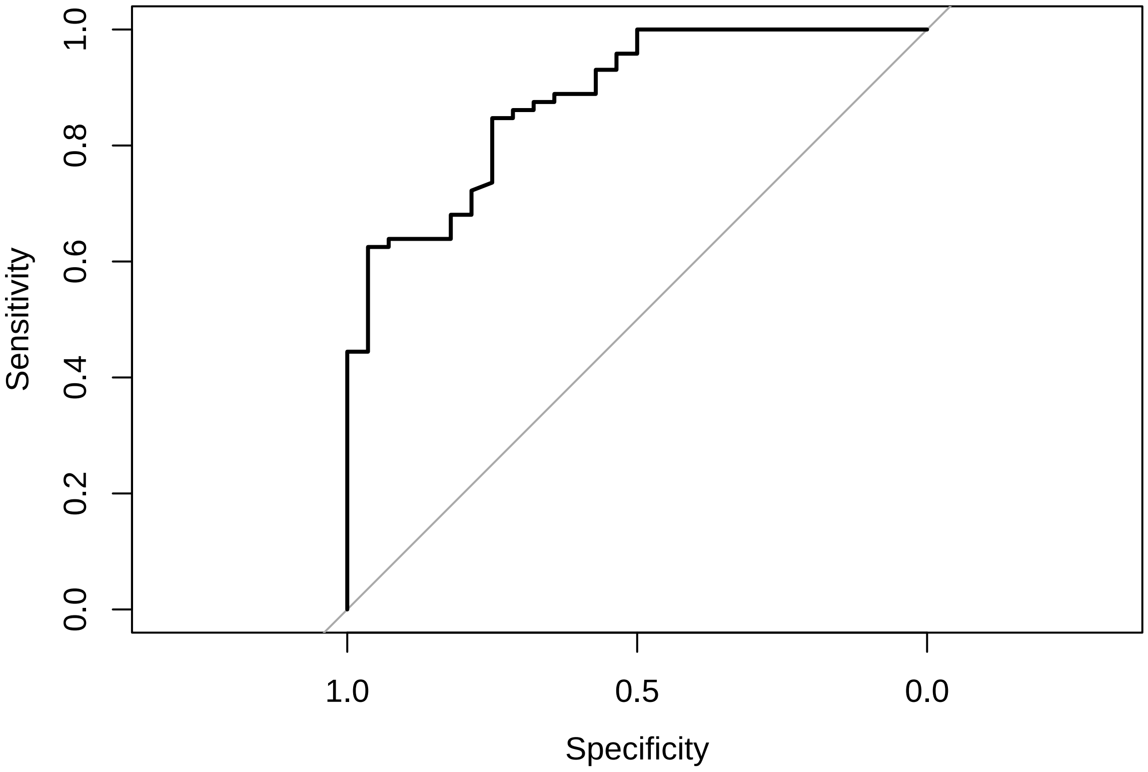
```
                  Estim. Low.lim(95%) Up.lim(95%)
Sensitivity    0.8194444    0.7074138   0.8966915
Specificity    0.7500000    0.5477916   0.8857060
Pos.Pred.Val.  0.8939394    0.7877070   0.9527189
Neg.Pred.Val.  0.6176471    0.4362474   0.7731278
LR+            3.2777778    1.7100443   6.2827771
LR-            0.2407407    0.1407775   0.4116858
Odds ratio    13.6153846    4.7877602  38.7192946
Youden index   0.5694444    0.3860923   0.7527966
Accuracy       0.8000000    0.7056770   0.8707518
Error rate     0.2000000    0.1292482   0.2943230
```

Now let's compute the ROC curve and the area under the curve (AUC) for a global assessment of the continuous marker diagnostic ability. For this aim we need to load the package *pROC* and use the function *roc(response, predictor,levels,plot)*. The arguments are:

- *response.* Outcome variable.

- *predictor*. Marker values.
- *levels*. The value of the response for non-diseased and diseased respectively.
- *plot*. plot the ROC curve? TRUE or FALSE

```
library(pROC)
roc(dades$Outcome, dades$Test1,levels=c("NI","I"),plot=T)
```



```
Call:
roc.default(response = dades$Outcome, predictor = dades$Test1,     levels = c("NI", "I")

Data: dades$Test1 in 28 controls (dades$Outcome NI) < 72 cases (dades$Outcome I).
Area under the curve: 0.8802
```

The value of the AUC indicates that the diagnostic ability is good.