# Introduction to probability computation and applications

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

# Probability

## Definition

- The probability is a measure of likelihood that an event occurs.
- It takes values between 0 and 1.
- A value of 0 means that is completely unlikely to observe the event whilst a value of 1 implies that the event is undoubtedly going to happen.
- In case the probability is 0.5 it is equally possible that the event happens as it does not.

## Some events

- **Elementary event**. Such an event that can not be divided in other events.
- **Compound event**. An event composed of elementary or other compound events.
- Example. Roll one dice.

  - Elementary event: "The result is 7". It only involves the result "7".

  The probability is: $P\left("7"\right) = \frac{1}{6}$
  where the value of 6 in the denominator stands for all the possible results.

  - Compound event: "The result is even". It involves the set of elementary results "1","3","5".

  The probability is: $P\left("even"\right) = \frac{3}{6} = \frac{1}{2}$

- **Intersection**. Two events A and B happen at the same time.

If the two events are independent $P\left(A\cap B\right)=P\left(A\right)P\left(B\right)$

- Example. Roll two dice. What is the probability of obtaining two ones? The results of each die can be assumed to be independent.

$$P\left(D_1=1\cap D_2=1\right)=P\left(D_1=1\right)P\left(D_2=1\right)=\frac{1}{6}\frac{1}{6}=\frac{1}{36}$$

- **Union**. Given two events A and B, at least one of them is observed.

$$P\left(A\cup B\right)=P\left(A\right)+P\left(B\right)-P\left(A\cap B\right)$$

- Example. Roll two dice ($D_1$ and $D_2$). What is the probability of obtaining one five at least? The results of each die can be assumed to be independent.

$$P\left(D_1=5\cup D_2=5\right)=P\left(D_1=5\right)+P\left(D_2=5\right)-P\left(D_1=5\cap D_2=5\right)=\frac{1}{6}+\frac{1}{6}-\frac{1}{36}=\frac{11}{36}$$

- **Complementary / contrary**. The opposite of an event. $A^C$

$$P\left(A^C\right)=1-P\left(A\right)$$

- Example. When rolling a die, what is the probability of obtaining a result different to “5”?

$$P\left(D\neq 5\right)=1-P\left(D=5\right)=1-\frac{1}{6}=\frac{5}{6}$$

## Conditional probability

- The probability of event A occurring if event B occurred.

$$P\left(A|B\right)$$

- If $P\left(A|B\right)=P\left(A\right)$, A and B are **independent**, i.e. the probability of A is not conditioned by B.

- Example. Roll a dice. What is the probability of “6” if we know that the result is even?

Two ways to find the probability.

a) Use the equation for conditional probability.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(D = 6|even) = \frac{P(D = 6 \cap even)}{P(even)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

b) Reduce the sample space (possible results) and compute the probability in this new sample space.

If we know that the result is even, the sample space is $\{2, 4, 6\}$. In this new sample space the probability of “6” is $\frac{1}{3}$.

Both approaches are equivalent.

- **Conditional complimentary**. The complimentary of a conditional event: keep the condition and apply the contrary to the event.

$$P\left((A|B)^C\right) = P\left(A^C|B\right)$$

- Example. Roll a dice. What is the probability of result different to “6” if we know that the result is even?

$$P\left((D = 6|\text{"}even\text{"})^C\right) = P(D \neq 6|\text{"}even\text{"})$$

a) Use the equation for conditional probability

$$P(D \neq 6|even) = \frac{P(D \neq 6 \cap even)}{P(even)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$

b) Reduce the sample space (possible results) and compute the probability in this new sample space.

If we know that the result is even, the sample space is $\{2, 4, 6\}$. In this new sample space the probability of not “6” is $\frac{2}{3}$.

**Bayes theorem**

- Suppose the information is brought to us in terms of conditional probabilities rather than intersections of elementary events.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

but also

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

so that

$$P(A \cap B) = P(B|A)\, P(A)$$

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}$$

In this case the probability of B has to be computed considering all the settings in relation of A.

$$P(B) = P(A \cap B) + P\left(A^C \cap B\right) = P(B|A)\, P(A) + P\left(B|A^C\right) P\left(A^C\right)$$

Example. It is known that 1 out of 3 subjects using a treatment have a positive response. Among them 1 out of 4 have genotype AA whereas only a 10% of subjects with negative response have genotype AA.

$$P(+) = \frac{1}{3}; P(AA|+) = \frac{1}{4}; P(AA|-) = \frac{1}{10}$$

What is the probability of positive response if a subject has genotype AA?

$$P(+|AA) = \frac{P(AA|+)\, P(+)}{P(AA|+)\, P(+) + P(AA|-)\, P(-)} = $$

$$= \frac{\frac{1}{4}\frac{1}{3}}{\frac{1}{4}\frac{1}{3} + \frac{1}{10}\frac{2}{3}} = \frac{5}{9}$$

# Assessment of diagnostic test ability

## Introduction

- Marker values are used as a predictor of a binary outcome.
- **Diagnostic test**. The marker values and the outcome are referred to the same chronological time.
- **Prognostic test**. The marker values are obtained before the outcome happens.

Examples: clinical signs or symptoms, biomarkers, laboratory tests,...

- Commonly the marker is dichotomized as a positive or negative result.

If the test ($Y = 0, 1$) and the outcome ($D = 0, 1$) are binary variables, there are four possible intersections between the two variables.

| | Outcome | | |
|---|---|---|---|
| Test | $D = 1$ | $D = 0$ | |
| $Y = 1$ | TP | FP | $TP + FP$ |
| $Y = 0$ | FN | TN | $FN + TN$ |
| | $TP + FN$ | $FP + TN$ | |

- $D = 1$ for outcome present. $D = 0$ for outcome absent.
- $Y = 1$ for a positive test. $Y = 0$ for a negative test.
- TP: True positive. $(Y = 1) \cap (D = 1)$
- FP: False positive. $(Y = 1) \cap (D = 0)$
- TN: True negative. $(Y = 0) \cap (D = 0)$
- FN: False negative. $(Y = 0) \cap (D = 1)$

**Example**. Ultrasonography approach is assessed to diagnose prostate cancer. A sample of 245 subjects, 105 with prostate cancer and 140 controls, were undergo to an ultrasonography. The results are in the following table:

| | Outcome | | |
|---|---|---|---|
| Test | $D = 1$ | $D = 0$ | |
| $Y = 1$ | 45 | 50 | 95 |
| $Y = 0$ | 60 | 90 | 150 |
| | 105 | 140 | 245 |

## Accuracy measures

- **Accuracy**: the test's ability to correctly detect a condition when it is actually present and to rule out a condition when it is truly absent.
- Accuracy has to do with concepts as:
  - Classification error
  - Association between test and outcome
  - Discriminative ability of outcome populations

### Sensitivity and Specificity

- True positive fraction (TPF)

$$TPF = P\left(Y = 1|D = 1\right)$$

In Medicine is known as **Sensitivity**.

**Sensitivity**: It assesses the test's ability to detect the condition when it is present.

In Engineering is also known as **Success rate**

- False positive fraction (FPF)

$$FPF = P\left(Y = 1|D = 0\right)$$

In Medicine the complementary of FPF, $(1 - FPF)$, is known as **Specificity**.

**Specificity**: It assesses the test's ability to exclude the condition in subjects without the condition.

In Engineering FPF is known as **False alarm rate**

In the example:

$$TPF = \frac{45}{105} = 0.429$$
$$FPF = \frac{50}{140} = 0.357$$
$$\text{Specificity} = 1 - FPF = 0.643$$

**Odds Ratio**

It is a measure of association between the test and the outcome.

First let's define an odds.

An **odds** is the ratio between a probability and its complementary $\frac{p}{1-p}$.

- Odds of outcome present.

$$Odds_{D=1} = \frac{P\left(Y=1|D=1\right)}{P\left(Y=0|D=1\right)} = \frac{TPF}{1-TPF}$$

- Odds of outcome absent.

$$Odds_{D=0} = \frac{P\left(Y=1|D=0\right)}{P\left(Y=0|D=0\right)} = \frac{FPF}{1-FPF}$$

- Interpreting the Odds
  - $Odds > 1$ means a positive result is **more probable** than a negative one.
  - $Odds = 1$ means a positive result is **equally probable** than a negative one.
  - $Odds < 1$ means a positive result is **less probable** than a negative one.

If the test performs well it is expected that:

- $Odds_{D=1}$ takes values greater than 1.
- $Odds_{D=0}$ takes values lower than 1.

**Odds ratio**. Ratio of two odds.

$$OR_{\frac{D=1}{D=0}} = \frac{Odds_{D=1}}{Odds_{D=0}}$$

- $OR > 1$. The Odds of $D = 1$ is greater than Odds for $D = 0$. The test performs correctly.
- $OR = 1$. The Odds of $D = 1$ is equal than Odds for $D = 0$. The test is useless. Independence between test and outcome.
- $OR < 1$. The Odds of $D = 1$ is lower than Odds for $D = 0$. Nonsense. Revise the test.

In the **example**:

$$Odds_{D=1} = \frac{TPF}{1-TPF} = \frac{45}{60} = 0.75$$
$$Odds_{D=0} = \frac{FPF}{1-FPF} = \frac{50}{90} = 0.556$$
$$OR = \frac{0.75}{0.556} = 1.35$$

**Likelihood Ratios**

Likelihood ratios are interpreted as "information gain". How much knowledge do the test bring about the outcome?

- **Positive likelihood ratio**.

$$LR+ = \frac{P\left(Y=1|D=1\right)}{P\left(Y=1|D=0\right)} = \frac{TPF}{FPF}$$

How much probable is a positive result when outcome is present.

- **Negative likelihood ratio**.

$$LR- = \frac{P\left(Y=0|D=1\right)}{P\left(Y=0|D=0\right)} = \frac{1-TPF}{1-FPF}$$

How much probable is a negative result when outcome is present.

- **Pre-test Odds**

$$Odds_{Pre} = \frac{P\left(D=1\right)}{P\left(D=0\right)}$$

- **Post-test Odds**

$$Odds_{Post,Y=1} = \frac{P\left(D=1|Y=1\right)}{P\left(D=0|Y=1\right)} = Odds_{Pre} \cdot LR+$$

$$Odds_{Post,Y=0} = \frac{P\left(D=1|Y=0\right)}{P\left(D=0|Y=0\right)} = Odds_{Pre} \cdot LR-$$

Likelihood ratios informs how the test modifies the odds.

- $LR+$ indicates how much informative is a positive result.
- $LR-$ indicates how much informative is a negative result.

In the example:

$$LR+ = \frac{TPF}{FPF} = 1.2$$
$$LR- = \frac{1-TPF}{1-FPF} = 0.89$$

## Utility measures

- How much useful is the test when applied?
- **Positive predictive value (PPV)**

Utility of positive results.

$$PPV = P\left(D = 1|Y = 1\right) = \frac{TPF \cdot p}{TPF \cdot p + FPF \cdot (1 - p)}$$

- **Negative predictive value (NPV)**

Utility of negative results.

$$NPV = P\left(D = 0|Y = 0\right) = \frac{(1 - FPF) \cdot (1 - p)}{(1 - FPF) \cdot (1 - p) + (1 - TPF) \cdot p}$$

In the example:

- If the prevalence of the table is representative of the "true prevalence"

$$p = P\left(D = 1\right) = \frac{105}{245} = 0.43$$

$$PPV = P\left(D = 1|Y = 1\right) = \frac{45}{95} = 0.474$$

$$NPV = P\left(D = 0|Y = 0\right) = \frac{60}{150} = 0.6$$

- Utility if the prevalence is 0.2.

$$PPV = P\left(D = 1|Y = 1\right) = \frac{0.429 \cdot 0.2}{0.429 \cdot 0.2 + 0.357 \cdot 0.8} = 0.231$$

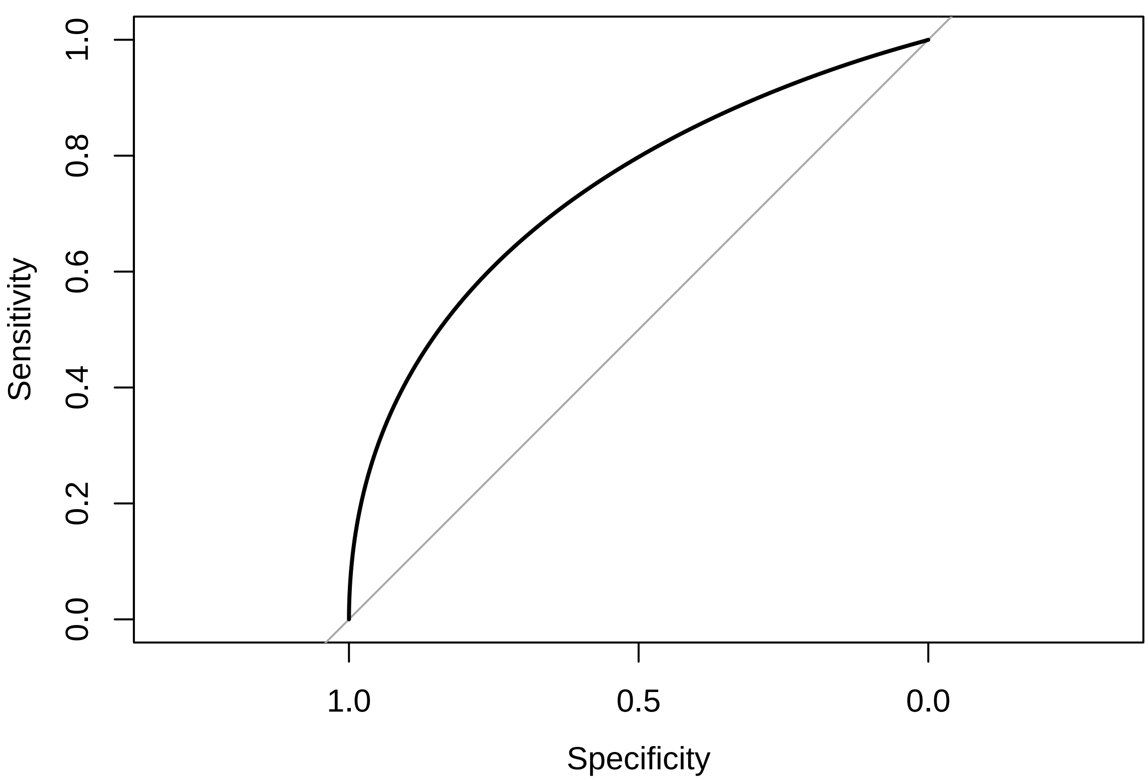
$$NPV = P\left(D = 0|Y = 0\right) = \frac{0.643 \cdot 0.8}{0.643 \cdot 0.8 + 0.571 \cdot 0.2} = 0.818$$

## Continuous test

- Let Y be a continuous variable.
- Usually the test is dichotomized using a threshold ($c$) and proceed as we saw before (accuracy and utility measures).
- If the population with outcome present ($D = 1$) has greater values (mean) than the population with outcome absent ($D = 0$)

$Y \geq c \rightarrow$ positive and $Y < c \rightarrow$ negative.

- Notice that accuracy and utility measures change with a different threshold.
- Plotting the pairs $(TPF, 1 - FPF)$ produces a **ROC curve**.
- Area under the ROC curve (AUC) is used to assess the diagnostic ability of the test.




```
Area under the curve: 0.74
```

- AUC takes values between 0.5 (independence between marker and outcome) and 1 (perfect discrimination).

- Interpretation of AUC:
  - $0.5 < AUC < 0.7$, diagnostic ability is poor.
  - $0.7 \leq AUC < 0.9$ and $< 0.9$ diagnostic ability is good.
  - AUC $\geq 0.9$, diagnostic ability is excellent.