Bayes' Theorem

For two events A and B, if we know the conditional probability $P(B|A)$ and the probability $P(A)$, then the Bayes' theorem tells that we can compute the conditional probability $P(A|B)$ as follows:

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

In statistics, the Bayes' theorem is often used in the following way:

$$P(\text{Unknown}|\text{Data}) = \frac{P(\text{Data}|\text{Unknown})P(\text{Unknown})}{P(\text{Data})}$$

Example: Accuracy of X-rays (continued)

Table 1: X-ray reading data

	Persons without TB	Persons with TB	Total
+ X-ray	51	22	73
− X-ray	1739	8	1747
Total	1790	30	1820

D_1: tuberculosis

D_2: no tuberculosis

T^+: positive X-ray

T^-: negative X-ray

It is useful to find $P(D_1|T^+)$, the probability that an individual has the disease given that he tests positive. This probability is also called the predictive value of a positive test.

Remark: Here because of the problem of sampling bias, it is not correct to simply estimate $P(D_1|T^+)$ based on the observed data, i.e., the numbers given in the table. This incorrect estimate gives 22/73, which has a large upward bias and over estimates $P(D_1|T^+)$.

We use the Bayes theorem to find $P(D_1|T^+)$. The Bayes theorem says:

$$P(D_1|T^+) = \frac{P(T^+|D_1)P(D_1)}{P(T^+)}.$$

There are two importants things here:

1. **Prior probability:** $P(D_1)$: the probability of TB before having the data. This is called prior probability. Usually, a judgement call has to be made as to what prior probability to use. For the present problem, it seems reasonable to use the population prevalence as the prior probability. In 1987, there were 9.3 TB cases per 100,000 population. Therefore, we specify:

$$P(D_1) = \frac{9.3}{100,000} = 0.000093.$$

2. **Probability of the data given the model (the likelihood):** $P(T^+|D_1)$: the probability of test positive given the disease. This summarizes the information in the data, i.e., the test results in our problem. Here

$$P(T^+|D_1) = \frac{22}{30} = 0.7333333$$

Then operationally, the Bayesian method is to make inference based on the **posterior probability**, calculated from the prior probability and the **probability of the data given the model (likelihood)** using the Bayes theorem.

We can calculate $P(T^+)$ as follows:

$$\begin{aligned} P(T^+) &= P(T^+ \cup D_1) + P(T^+ \cup \overline{D}_1) \\ &= P(T^+|D_1)P(D_1) + P(T^+|\overline{D}_1)P(\overline{D}_1). \end{aligned}$$

Plug in the numbers, we get

$$P(T^+) = \frac{22}{30} \times 0.000093 + \frac{51}{1790} \times (1 - 0.000093) = 0.02855717.$$

Therefore,

$$P(D_1|T^+) = \frac{0.7333333 \times 0.000093}{0.02855717} \approx 0.00239.$$

We note that

$$\frac{P(D_1|T^+)}{P(D_1)} = \frac{0.00239}{0.000093} = 25.7.$$

So the probability that an individual with a positive X-ray has TB is about 26 times greater than the probability for an individual randomly chosen from the population.

The relative risk and the odds ratio

Relative risk

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

Here the exposure can be either environmental or genetic (or both).

$$P(\mathsf{death\ over\ 35\ due\ to\ lung\ cancer}|\mathsf{male\ smoker}) = 0.002679$$
$$P(\mathsf{death\ over\ 35\ due\ to\ lung\ cancer}|\mathsf{male\ nonsmoker}) = 0.000154$$

$$\mathsf{RR} = \frac{0.002679}{0.000154} = 17.4.$$

In **genetic epidemiology**, the *genotypic relative risk* (GRR) is an important quantity to measure the genetic contribution to a disease. The GRR also gives indication on how difficult it is to identify the chromosomal regions that may harbor the disease-predisposing genes.

Suppose the genotypes at a disease-predisposing locus are AA, Aa and aa, where the allele A increases the risk of disease. Then

$$\text{GRR}_1 = \frac{P(\text{disease}|Aa)}{P(\text{disease}|aa)}$$
$$\text{GRR}_2 = \frac{P(\text{disease}|AA)}{P(\text{disease}|aa)}$$

For the Mendelian diseases (e.g., Cystic Fibrosis, Huntington's disease), the GGR is very high. However, for many common diseases, although there usually is a strong genetic component that contributes to elevating the risk of the disease, the GRR is often relatively low. This makes it difficult to map the genes that predispose the disease (e.g. alcoholism, autism, bipolar).

Odds ratio

Odds

Let

$$p_1 = P(\text{disease}|\text{exposed}).$$

So p_1 is the probability that an individual has the disease given that he/she is exposed to a certain risk factor. The odds of having the disease given exposure to the risk factor is

$$\text{odds.disease(exposed)} = \frac{p_1}{1-p_1}.$$

Likewise, let p_2 be the probability that an individual has the disease given that he/she is not exposed to a certain risk factor. The odds of having the disease given no exposure to the risk factor is

$$\text{odds.disease(unexposed)} = \frac{p_2}{1-p_2}.$$

Odds ratio

$$\begin{aligned}\text{OR} &= \frac{\text{odds.disease(exposed)}}{\text{odds.disease(unexposed)}} \\ &= \frac{P(\text{disease}|\text{exposed})/[1-P(\text{disease}|\text{exposed})]}{P(\text{disease}|\text{unexposed})/[1-P(\text{disease}|\text{unexposed})]}.\end{aligned}$$

The OR can also be defined as the ratio of the odds of exposure among diseased individuals and the odds of exposure among nondiseased individuals.

$$\text{OR} = \frac{P(\text{exposure}|\text{diseased})/[1-P(\text{exposure})|\text{diseased})]}{P(\text{exposure}|\text{nondiseased})/[1-P(\text{exposure}|\text{nondiseased})]}.$$

The above two definitions are mathematically equivalent. The second definition is useful in $case\text{-}control$ studies.

Proof of the equivalence of the two expressions for OR

Let $D = \{\text{diseased}\}, \bar{D} = \{\text{nondiseased}\}$
and $E = \{\text{exposed}\}, \bar{E} = \{\text{unexposed}\}$.
Because

$$\begin{aligned}
1 - P(\text{disease}|\text{exposed}) &= 1 - P(D|E) = P(\bar{D}|E) \\
1 - P(\text{disease}|\text{unexposed}) &= 1 - P(D|\bar{E}) = P(\bar{D}|\bar{E}) \\
1 - P(\text{exposure})|\text{diseased}) &= 1 - P(E|D) = P(\bar{E}|D) \\
1 - P(\text{exposure}|\text{nondiseased}) &= 1 - P(E|\bar{D}) = P(\bar{E}|\bar{D}),
\end{aligned}$$

what we need to prove is

$$\frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(D|\bar{E})} = \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})}.$$

By the Bayes' theorem, the left-hand side equals

$$\begin{aligned}
\frac{P(D|E)/P(\bar{D}|E)}{P(D|\bar{E})/P(D|\bar{E})} &= \frac{\frac{P(E|D)P(D)}{P(E)} / \frac{P(E|\bar{D})P(\bar{D})}{P(E)}}{\frac{P(\bar{E}|D)P(D)}{P(\bar{E})} / \frac{P(\bar{E}|D)P(D)}{P(\bar{E})}} \\
&= \frac{P(E|D)/P(E|\bar{D})}{P(\bar{E}|D)/P(\bar{E}|\bar{D})} \\
&= \frac{P(E|D)/P(\bar{E}|D)}{P(E|\bar{D})/P(\bar{E}|\bar{D})} \\
&= \text{Right-hand side.}
\end{aligned}$$

This proves that the two expressions for OR are equal.

Example: In a case control study, investigators start by identifying individuals with the disease (the cases) and without the disease (the controls). They then go back in time to determine whether the exposure is question was present or absent for each individual. In a study that examines the effects of the use of oral contraceptives on the breast cancer, among the 989 women who had breast cancer, 273 had previously used oral contraceptives and 716 had not. Of the 9901 women who did not have breast cancer, 2641 had used oral contraceptives and 7269 had not.

In such a study, the proportions of individuals with and without the disease are chosen by the investigator, therefore, the probability of disease in the exposed and unexposed groups cannot be estimated. However, we can estimate the probability of exposure for both cases and controls. Thus by the second form of the OR, we can calculate the OR of the disease of exposed verses unexposed.

Table 2: Breast cancer example

	Breast cancer		
Oral contraceptive	Yes	No	Total
Yes	273	2641	2914
No	716	7260	7976
Total	989	9901	10890

$$
\begin{aligned}
\mathsf{OR} &= \frac{(273/989)/(1-273/989)}{(2641/9901)/(1-2641/9901)} \\
&= \frac{(273/989)/(716/989)}{(2641/9901)/(7260/9901)} \\
&= \frac{273/716}{2641/7260} \\
&= \frac{273 \times 7260}{716 \times 2641} \\
&= 1.05.
\end{aligned}
$$

Relationship between OR and RR

If $P(\text{disease}|\text{exposed}) \approx 0$ and $P(\text{disease}|\text{unexposed}) \approx 0$, then

$$\begin{aligned}\text{OR} &= \frac{\text{odds.disease(exposed)}}{\text{odds.disease(unexposed)}} \\ &\approx \frac{P(\text{disease}|\text{exposed})}{P(\text{disease}|\text{unexposed})} \\ &= \text{RR}.\end{aligned}$$

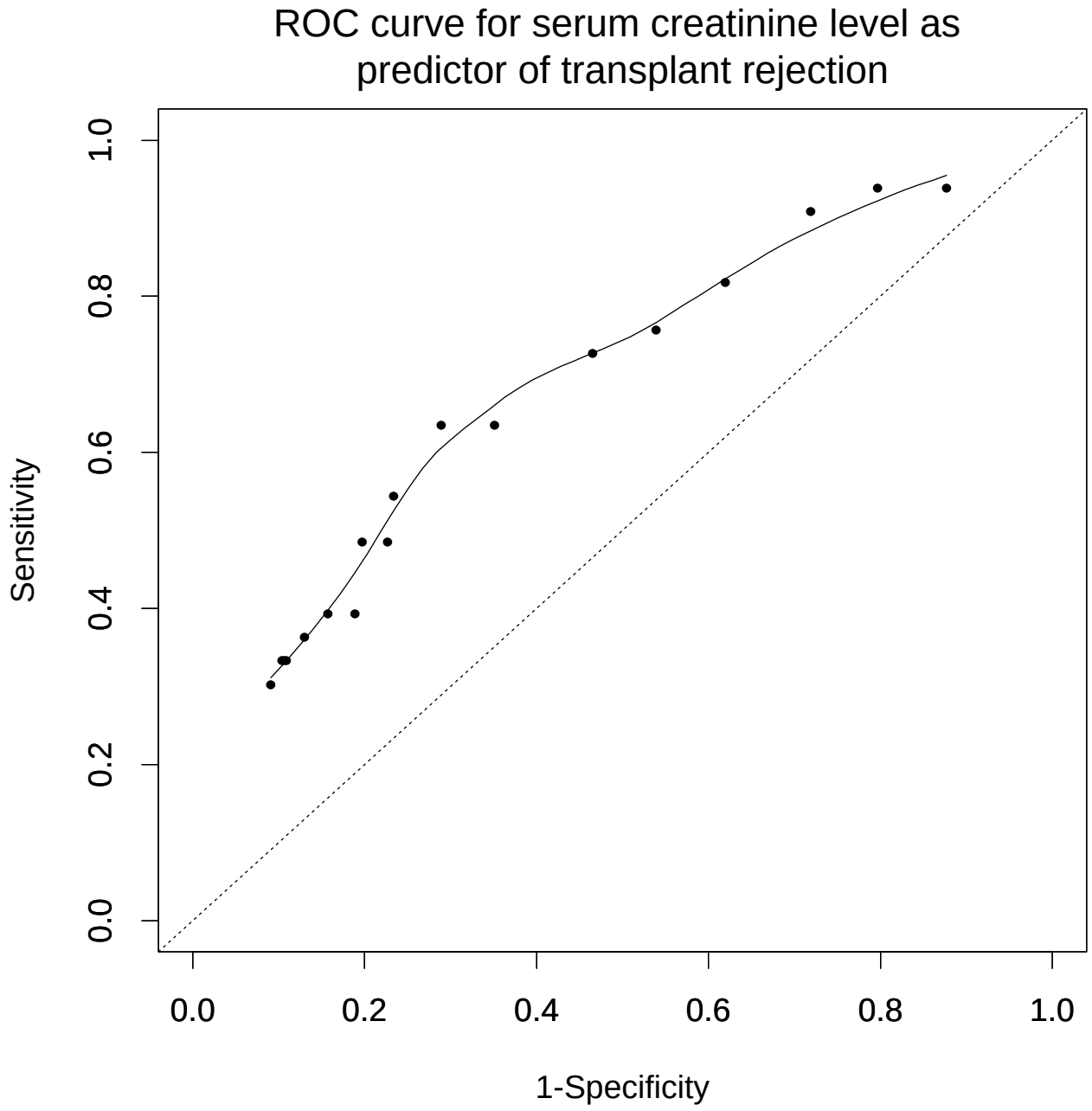
Receive Operator Characteristic (ROC) Curves

- Sensitivity: the probability of a positive test given that 'it' is present
- Specificity: the probability of a negative test given that 'it' is not present

The purpose of the ROC analysis is to find the trade-off (usually a threshold value) so that the levels of sensitivity and specificity are acceptable.

Table 3: Sensitivity and specificity of serum creatinine level for predicting transplant rejection

Serum Creatinine (mg %)	Sensitivity	Specificity
1.2	0.939	0.13
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	0.363	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


ROC curve for serum creatinine level as predictor of transplant rejection
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0