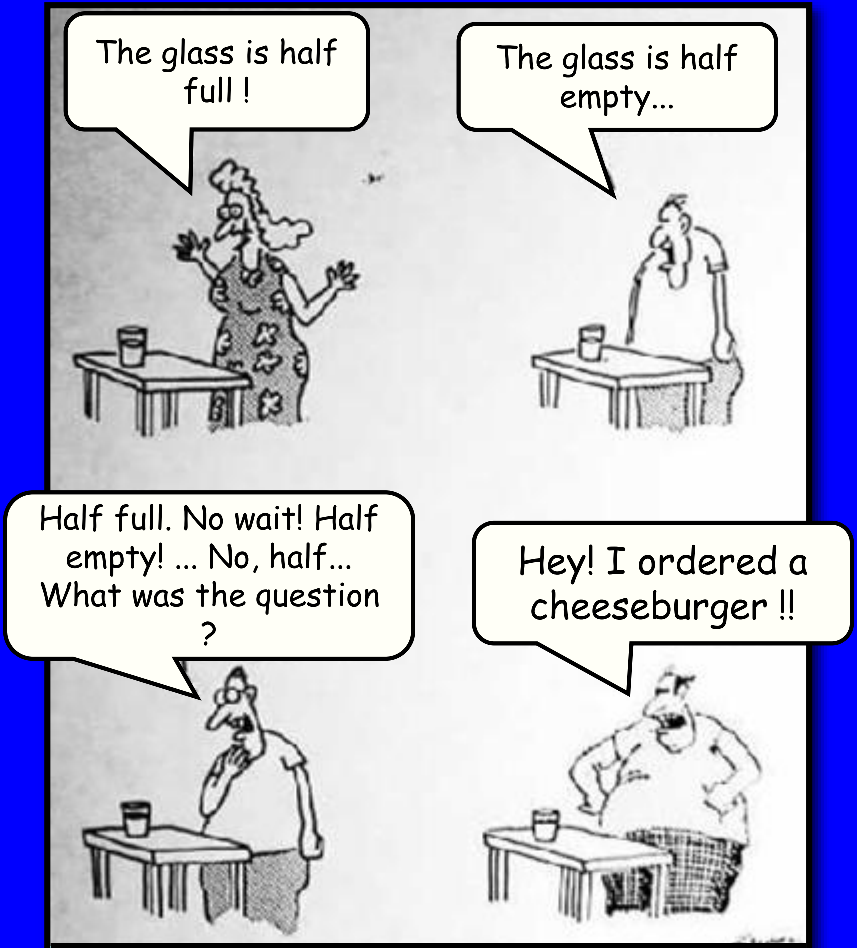
Lecture 8. Classification and discriminant analysis


The glass is half full !
The glass is half empty...
Half full. No wait! Half empty! ... No, half... What was the question ?
Hey! I ordered a cheeseburger !!

The four basic personality types

Overview

1. Classification
2. One dimension
3. Multiple dimensions (DA)
4. How good is solution?
5. Generalisation to other populations
6. Bayes' theorem

1. Classification

Dimensional versus categorical

- Psychometrics until now was mostly *dimensional:* assigning scores to people on (latent) dimensions (INT).
- *Decisions* based on these scores are mostly *categorical:* to which *group* must we allocate the person (NOM)?

Classification. Allocating individuals to groups (categories).

Examples in psychology

- *Clinical:* psychiatric diagnostics.
- *Educational:* advice on school choice based on primary school-leaving exam (CITO) scores.
- *Organisational:* personnel selection (partly) on basis of test scores.

General aim (in data language)

To predict categorical dependent variable Y (which distinguishes k groups from each other) as accurately as possible on the basis of p independent interval variables (X_1, X_2, ..., X_p).

- *One dimension* ($p = 1$): choose optimal cut-off point.
- *Multiple dimensions* ($p \geq 2$): discriminant analysis.

Terminology

- *X-variables:* dimensions, predictors, interval variables.
- *Y:* (group) classification, categorical variable, diagnosis.

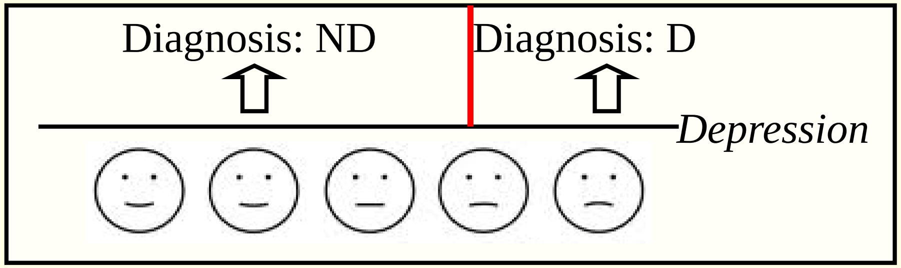
General procedure

1. Collect data about *X*-variables in a sample *where Y (classification) is already known*.
2. Look for *optimal prediction rule* to predict *Y* from the *X* variables as accurately as possible within sample. How good is this prediction?
3. Use this prediction rule for *new cases*, *where Y is not yet known*. How good is prediction in this new situation (different population)?

2. One dimension

Most simple case: two groups on one dimension

Example. Administer depression inventory to clinically depressed group (Y = D) and non-depressed controls (Y = ND).



Problem. Look for optimal *cut-off point* X_C such that:

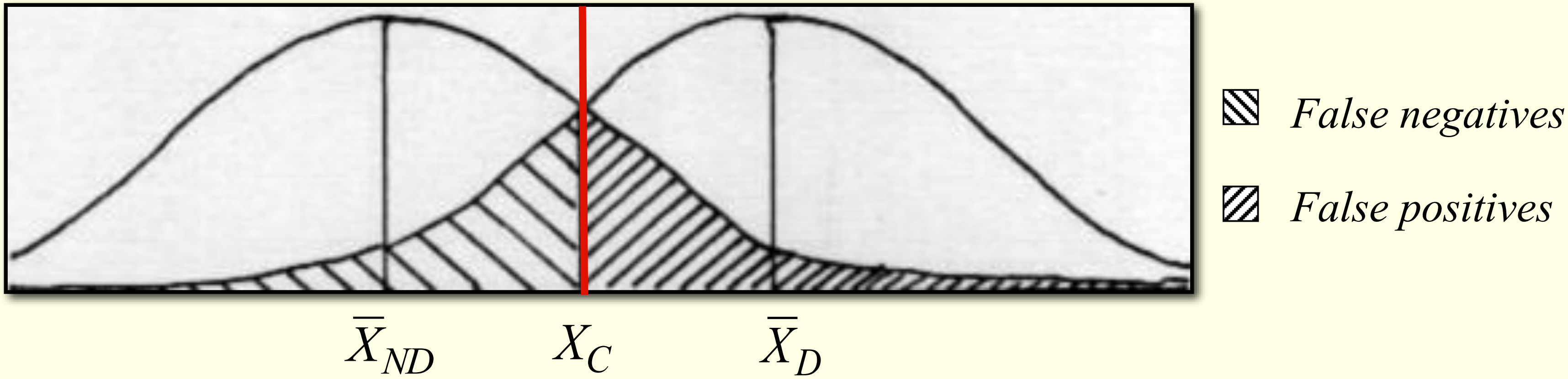
- if $X \geq X_C$: $\hat{Y}$ = D (→ predict D group);
- if $X < X_C$: $\hat{Y}$ = ND (→ predict ND group).

But what is "optimal"? Groups are rarely perfectly distinct from each other ⇒ always *errors* (allocating person to wrong group).

Two types of errors

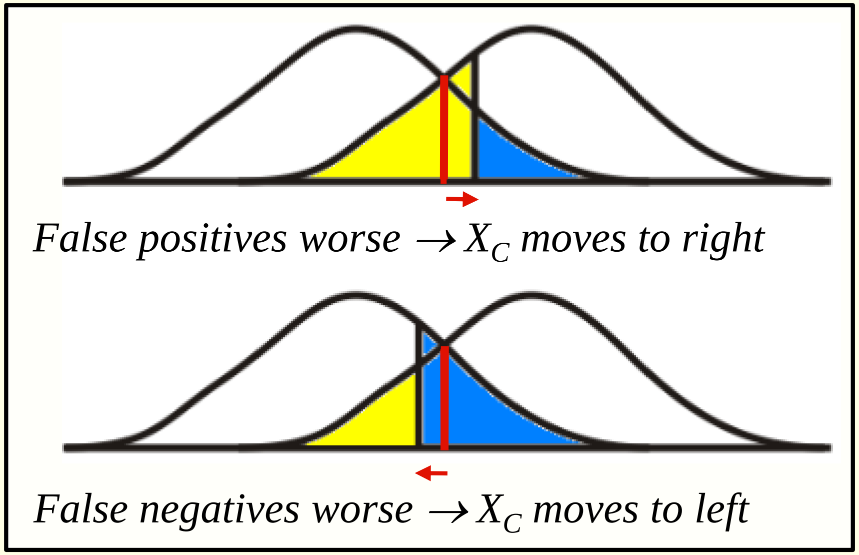
- *False positives.* Depression predicted (X = D) for actual controls (Y = ND).
- *False negatives.* Control predicted (X = ND) for actual depressives (Y = D).

		Prediction (X)	
		D	*ND*
Actual (Y)	*D*	True positives	False negatives
	ND	False positives	True negatives



Optimal = as few problems due to errors as possible. *"Problems due to"* errors are partly determined by *how bad* we regard different types of errors.

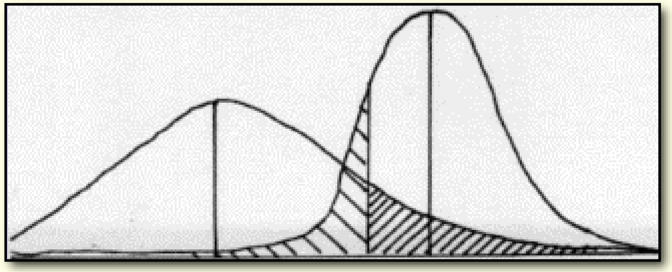
- *Both errors equally bad.* Minimise the sum of false positives (P) and false negatives (N) → $P + N$ as small as possible.
- *Some errors worse than others.* Minimise *weighted* sum of errors ($w_1P + w_2N$).
  - E.g. $2P + N$ (*false positives worse*): X_C moves to right → fewer false positives, but more false negatives.
  - Or $P + 2N$ *(false negatives worse):* now X_C moves to left.



False positives worse → X_C *moves to right*

False negatives worse → X_C *moves to left*

Even if both errors are equally bad, X_C does not always lie exactly between group means, e.g. different variances.



In short, even with just one predictor, optimal allocation of people to groups is not simple.

3. Multiple dimensions (DA)

Techniques for two or more dimensions (interval predictors):

- $p \geq 2$, $k = 2$: *logistic regression analysis (LRA)* or discriminant analysis (DA). With two groups, LRA is usually preferred.
- $p \geq 2$, $k > 2$: *discriminant analysis (DA).*

This course. DA is briefly discussed. LRA (and more DA) in MVDA (second semester).

Two sides to DA

- *Descriptive:* multivariate description of differences between groups. Not covered here, but in MVDA course.
  → A lot of SPSS output (e.g. about discriminant functions and their weights) not yet discussed.
- *Predictive*: individual prediction, allocation of individuals to groups (classification).

Three key problems

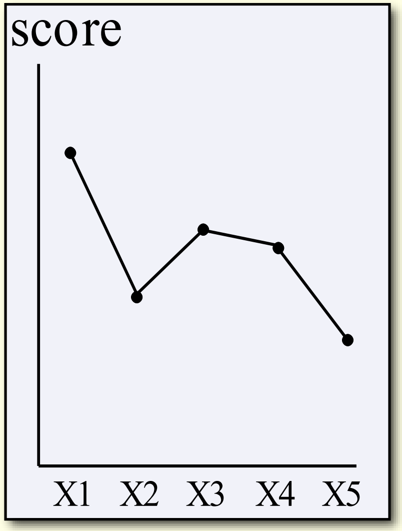
a. How to combine multiple dimensions in given sample for *optimal allocation* to groups?

b. *How good* is this optimal classification?

c. How to *generalise* the results *to groups other* than the original sample?

Classification within sample

- *Profile:* pattern of scores by individual or group on series of *p* variables.
- Imagine both individual and group profiles as *points* in a *p*-dimensional Euclidian *space* of variables.

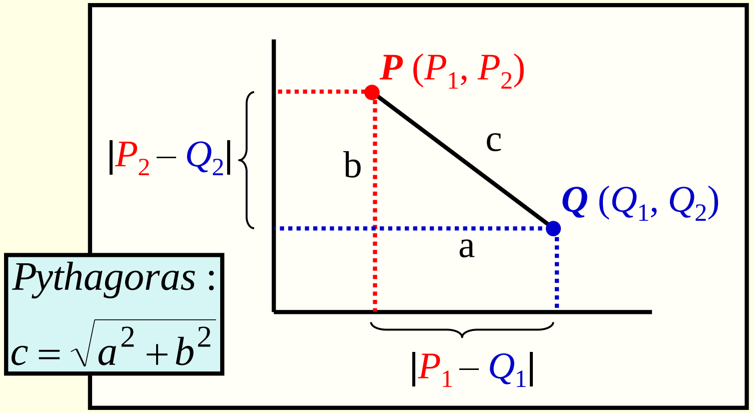


Allocation to groups: two steps

- For each individual, calculate distance from all group points (centroids) with *(generalised) Pythagorean theorem:*

$$d_{PQ} = \sqrt{\sum_{i=1}^{p} (P_i - Q_i)^2}$$

- Allocate individual to group with the shortest distance.



Example (2 variables, 2 groups)

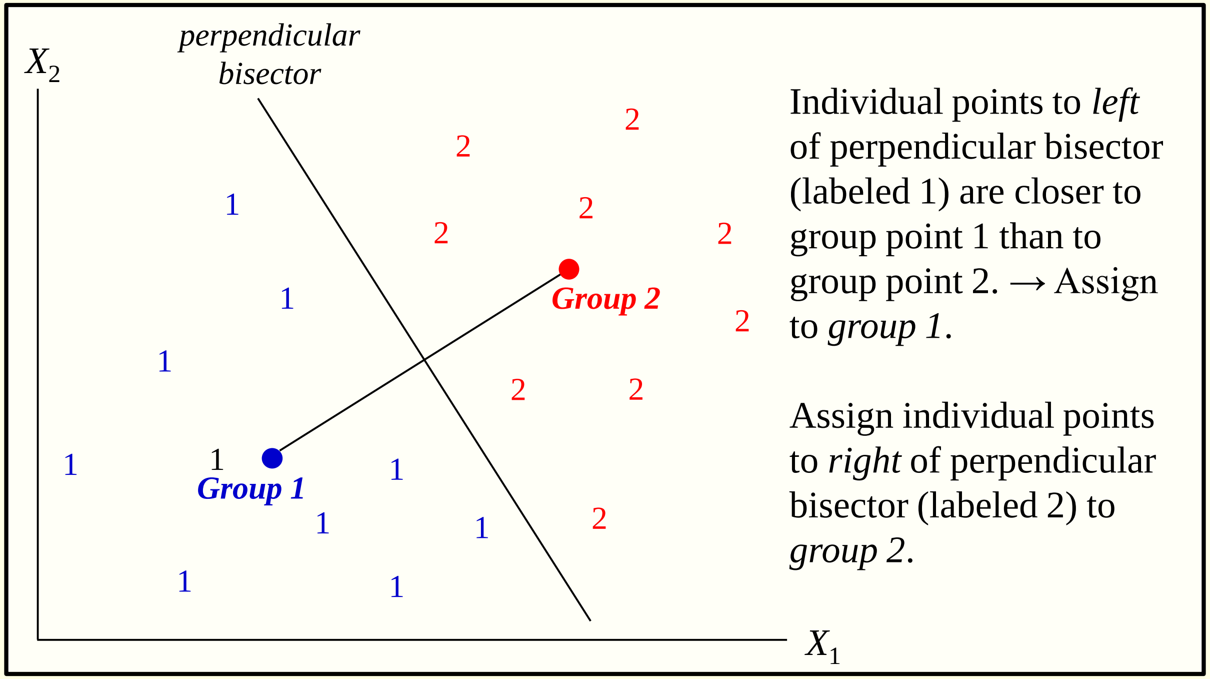
Example is too simple in many respects.

- Often *more than two dimensions* (variables).

  No real problem, because of generalised Pythagorean theorem.

- Complications due to *different variances* of:

  - *variables* → variables with relatively high variances more influential than other variables;
  - *groups* on the same variable → cut-off lines move toward groups with smallest variances.



Individual points to *left* of perpendicular bisector (labeled 1) are closer to group point 1 than to group point 2. → Assign to *group 1*.

Assign individual points to *right* of perpendicular bisector (labeled 2) to *group 2*.

- Variables can be *correlated*.
- Data can display *non-linear patterns*.

Variety of solutions (see syllabus). No problem for us, because SPSS does the work.

4. How good is solution?

Classification table. Cross-tabulation of *predicted* values (*X*: predicted group classification / diagnosis according to DA) compared with *real* values (*Y*: actual group classification).

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	80	20	100
	ND	11	89	100
	Total	91	109	200

Different quality measures based on classification tabel:

- PAC
- sensitivity and specificity
- positive and negative predictive value.

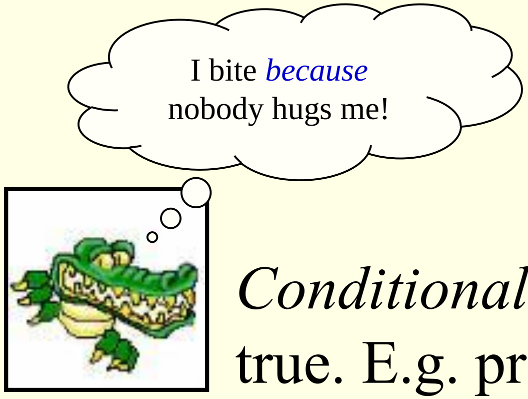
From classification table we can derive various measures.

1) Percentage accuracy in classification (PAC)

$$PAC = \frac{\textit{number of correct predictions}}{\textit{total number of predictions}} = \frac{80+89}{200} = \mathbf{.845}$$

PAC is very rough measure. Better to have measures based on *conditional probabilities*.

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	**80**	20	100
	ND	11	**89**	100
	Total	91	109	**200**



Conditional probability: p(B|H) = probability of *B* if *H* is true. E.g. probability of bite (*B*) ***if*** I hug crocodile (*H*).

Conditional probabilities important for *two kinds of questions*:

A. *Quality of measuring instrument.* Given an *actual situation* (Y_D or Y_{ND}), how high is probability of correct diagnosis?

B. *Quality of individual diagnosis.* Given a *diagnosis* (X_D or X_{ND}), how high is probability that this diagnosis is correct (how reliable is diagnosis, from recipient's viewpoint)?

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	**80**	20	**100**
	ND	11	89	100
	Total	91	109	200

A. Quality of measuring instrument

$$\textit{Sensitivity (D)} = \frac{\textit{number of correctly predicted D}}{\textit{total number of D}}$$

$$= p(X+|Y+) = 80 \,/\, 100 = \mathbf{.80}$$

probability of positive diagnosis if actually ill

$$\textit{Specificity (D)} = \frac{\textit{number of correctly predicted ND}}{\textit{total number of ND}}$$

$$= p(X\text{-}|Y\text{-}) = 89\ /\ 100 = \mathbf{.89}$$

probability of negative diagnosis if actually not ill

Sensitivity and specificity together determine quality of *measuring instrument.*

Ideal measuring instrument misses nobody who has disease (sensitivity = 1) and declares healthy everyone who does not have disease (specificity = 1).

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	80	20	100
	ND	11	**89**	**100**
	Total	91	109	200

Real measuring instruments will make *errors.* May lead to counterintuitive conclusions about quality of *individual diagnoses*.

B. Quality of individual diagnosis

Sensitivity and specificity say little about quality of individual diagnoses.

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	**80**	20	100
	ND	11	**89**	100
	Total	**91**	**109**	200

Individual wants different conditional probability: given a diagnosis (X_D or X_{ND}), probability that *I* actually belong to that group (*column* instead of row proportions).

$$\textit{Positive predictive value} = \frac{\textit{number of correctly predicted D}}{\textit{total predicted D}}$$

$$= p(Y+|X+) = 80 / 91 = \mathbf{.88}$$

probability that positive diagnosis is correct

$$\textit{Negative predictive value} = \frac{\textit{number of correctly predicted ND}}{\textit{total predicted ND}}$$

$$= p(Y-|X-) = 89 / 109 = \mathbf{.82}$$

probability that negative diagnosis is correct

5. Generalisation to other populations

With good samples (D and ND both representative of own subpopulation), sensitivity and specificity are *independent of proportions of* D and ND in investigated group.

This does not apply for positive and negative 'predictive value' !!!

Clinical group (100 D and 100 ND)

- Sensitivity and specificity .80 and .89.
- Positive and negative predictive value .88 and .82.

General population (e.g. 10000 people with 3% depressives)

If sensitivity and specificity remain the same (see above), we expect the following classification table.

How to create a classification table

1. Calculate *row totals* D and ND from size of population (N = 10000) and base rate (i.e. proportions of D and ND in population, e.g. N_{Y+} = .03 x 10000 = 300.

2a. Calculate number of *correctly diagnosed depressives* by multiplying total D by sensitivity: $N_{(X+\&Y+)}$ = .80 x 300 = 240.

2b. Calculate number of *correctly diagnosed non-depressives* by multiplying total ND by specificity: $N_{(X-\&Y-)}$ = .89 x 9700 = 8633.

3. Calculate numbers of *incorrect diagnoses* $N_{(X+\&Y-)}$ and $N_{(X-\&Y+)}$ by subtracting in each row number of correct diagnoses from row total, e.g. $N_{(X-\&Y+)}$ = 300 - 240.

4. Calculate *column totals* by adding together cell numbers in each column, e.g.

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	240 (=.80 x 300)	60 (=300-240)	300 (=.03 x 10000)
	ND	1067 (=9700-8633)	8633 (=.89 x 9700)	9700 (=.97 x 10000)
	Total	1307 (= 240+1067)	8693 (=60+8633)	10000

N_{X+} = 240 + 1067
= 1307.

		Diagnosis (X)		
		D	*ND*	*Total*
Actual (Y)	*D*	**240**	60	300
	ND	1067	**8633**	9700
	Total	**1307**	**8693**	10000

Predictive values become completely different:

- *PPV* = 240 / 1307 = **.184**
- *NPV* = 8633 / 8693 = **.993**

Only 18.4 % probability that someone with diagnosis Depression is actually depressed !!!

How is this possible?

Many more ND than D in population.

→ *Absolute numbers* of true and false positives are partly influenced by proportions of D and ND in population (“*base rates*”).

→ Many *more false positives than true positives*, because (1 - .89) x 9700 is much larger than .80 x 300.

Moral 1. Reliability of individual diagnoses is not only determined by quality of instruments, but also by '*base rate*' in population.

'Base rate' has no influence on sensitivity and specificity, but it does have an influence on numbers of true and false positives and negatives.

How good is prediction, all in all?

PAC = (240 + 8633) / 10000 = .89 Is that good?

Prediction *without* diagnostic information: assign everyone to the most frequent group (= best guess) → *everyone ND.*

→ PAC = 9700 / 10000 = .97.

Here, more correct predictions when we ignore diagnostic information (PAC = .97 versus .89), because:

- 240 (= 300 - 60) *more* false negatives;
- 1067 (= 1067 - 0) *less* false positives.

Moral 2. Sometimes, perhaps here too, it is better to ignore diagnostic information.

"Perhaps", because this partly depends on the relative seriousness of the two types of errors: missed diagnosis vs. false alarm.

6. Bayes' theorem

Calculations above are special case of *Bayes' theorem.*

General form

$$p(A|B) = \frac{p(A \,\&\, B)}{p(B)} = \frac{p(A \,\&\, B)}{p(A \,\&\, B) + p(\sim A \,\&\, B)}$$

$$= \frac{p(B \mid A)p(A)}{p(B \mid A)p(A) + p(B \mid \sim A)p(\sim A)}$$

In our case:
A = Y+ (having disease)
B = X+ (positive diagnosis)

Bayes for diagnoses

Makes it possible to derive *positive* and *negative predictive value*, i.e. $p(Y+|X+)$ and $p(Y-|X-)$, on the basis of:

- sensitivity and proportion of false positives (= 1 - specificity): $p(X+|Y+)$ and $p(X+|Y-)$, and
- base rates of positives and negatives in population: $p(Y+)$ and $p(Y-)$.

correct positive diagnosis

positive predictive value

probability of correct positive diagnosis

sensitivity

base rate positive

$$p(Y+\,|\,X+) \;=\; \frac{p(X+\&Y+)}{p(X+)} \;=\; \frac{p(X+\,|\,Y+)\,p(Y+)}{p(X+\,|\,Y+)\,p(Y+) \;+\; p(X+\,|\,Y-)\,p(Y-)}$$

total probability of positive diagnosis (correct or incorrect)

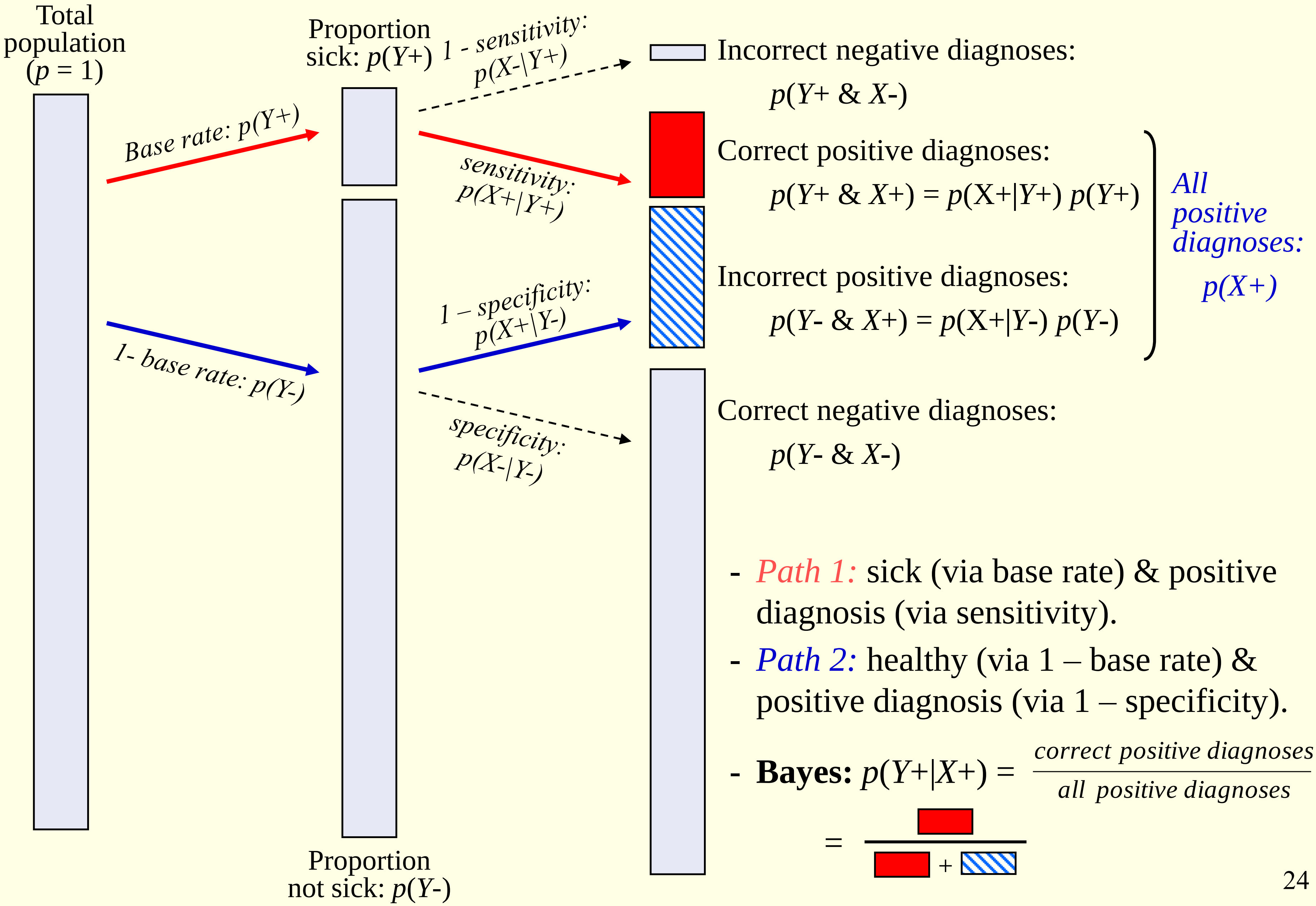
as numerator

1 - specificity

base rate negative

correct positive diagnosis ***incorrect positive diagnosis***

Graphical representation of Bayes' theorem



Total population (p = 1)

Base rate: p(Y+)

1- base rate: p(Y-)

Proportion sick: $p(Y+)$

Proportion not sick: $p(Y-)$

1 - sensitivity: p(X-|Y+)

sensitivity: p(X+|Y+)

1 – specificity: p(X+|Y-)

specificity: p(X-|Y-)

Incorrect negative diagnoses: $p(Y+ \;\&\; X-)$

Correct positive diagnoses: $p(Y+ \;\&\; X+) = p(X+|Y+)\, p(Y+)$

Incorrect positive diagnoses: $p(Y- \;\&\; X+) = p(X+|Y-)\, p(Y-)$

All positive diagnoses: $p(X+)$

Correct negative diagnoses: $p(Y- \;\&\; X-)$

- *Path 1:* sick (via base rate) & positive diagnosis (via sensitivity).
- *Path 2:* healthy (via 1 – base rate) & positive diagnosis (via 1 – specificity).
- **Bayes:** $p(Y+|X+) = \dfrac{\textit{correct positive diagnoses}}{\textit{all positive diagnoses}}$

= 🟥 / (🟥 + ▨)

Depression example

$$p(Y+ \mid X+) = \frac{.80 * .03}{.80 * .03 \; + .11 * .97} = \frac{.024}{.024 \; + .1067} = .184$$

Advantages of Bayes' theorem

- Makes connection between ad hoc solution and wider statistical theory.
- We do not need to know size of population (not relevant), because we can work directly with proportions.
- Generalizable to situations with more than two categories (e.g. unipolar depressive vs. bipolar depressive vs. non-depressive).