relationships between categorical variables

- Section 2.1 Basic terminology and notation.
- Parameters in Section 2.2 are used to compare groups on the proportions of responses in the outcome categories.
  - **Difference of Proportions**
  - **Relative Risk**
  - **odds ratio**
- In Section 2.3 we extend the scope by controlling for a third variable.
- The chapter's primary focus is binary variables, which have only two categories
- in Section 2.4 we present parameters for nominal and ordinal multicategory variables.

2.1 PROBABILITY STRUCTURE FOR CONTINGENCY TABLES

- The joint distribution between two categorical variables determines their relationship.
  - **Poisson**
  - **Binomial**
  - **Multinomial Sampling**
- This distribution also determines the marginal and conditional distributions.

2.1.1 Contingency Tables and Their Distributions

TABLE 2.1 Cross-Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

Source: Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New Engl. J. Med.* **318**: 262–264 (1988).

Table 2.1, a 2X3 contingency table, is from a report on the relationship between aspirin use and heart attacks

IxJ Contingency Table

- Let *X and Y denote two categorical response variables, X with I categories* and *Y with J categories.*
- *Classifications of subjects on both variables have IJ* possible combinations.
- The *cells of the table represent the IJ possible outcomes*
- When the cells contain frequency counts of outcomes for a sample, the table is called a *contingency table*
- Another name is *cross-classification table. A contingency table with I rows* and *J columns is called an IXJ or I-by-J table.*

Study: aspirin use and heart attacks by Harvard Medical School

- 5 year *randomized*, *blind* study
- Two arms, 1:1 ratio
  - Placebo
  - Aspirin
- Arm1 – placebo: of 11,034, 18 had fatal heart attacks
  Arm2 – aspirin: of 11,037, 5 had it

	Myocardial Infarction		
	Fatal Attack	Nonfatal Attack	No Attack
Placebo	18	171	10,845
Aspirin	5	99	10,933

SAS program

```
DATA AspirinStudy;
  input Treatment $ x y z; drop x y z;
  outcome='  Fatal Attack  '; w=x; output;
  outcome=' NonFatal Attack'; w=y; output;
  outcome='No Attack'; w=z; output;
  cards;
Placebo  18 171  10845
Aspirin  5   99  10933
;
proc freq data=AspirinStudy;
  weight w;
  table Treatment*outcome;
run;
```

Original data

ID	Race	Age	Treatment	Outcome
1	W	32	Placebo	No Attack
2	B	33	Aspirin	No Attack
…	…	…	…	…
…	…	…	…	…
22071	W	50	Aspirin	Fatal Attack

SAS data (aggregated data)

Treatment	outcome	w
Placebo	Fatal Attack	18
Placebo	NonFatal Attack	171
Placebo	No Attack	10845
Aspirin	Fatal Attack	5
Aspirin	NonFatal Attack	99
Aspirin	No Attack	10933

Table of Treatment by outcome

Treatment	outcome			
Frequency Percent Row Pct Col Pct	Fatal Attack	NonFatal Attack	No Attack	Total
Aspirin	5	99	10933	11037
	0.02	0.45	49.54	50.01
	0.05	0.90	99.06	
	21.74	36.67	50.20	
Placebo	18	171	10845	11034
	0.08	0.77	49.14	49.99
	0.16	1.55	98.29	
	78.26	63.33	49.80	
Total	23	270	21778	22071
	0.10	1.22	98.67	100.00

Table of Treatment by outcome

Treatment	outcome			
Frequency Percent Row Pct Col Pct	Fatal Attack	NonFatal Attack	No Attack	Total
Aspirin	n_{11} π_{11} $\pi_{1\|1}$	n_{12} π_{12} $\pi_{2\|1}$	n_{13} π_{13} $\pi_{3\|1}$	n_{1+} π_{1+}
Placebo	n_{21} π_{21} $\pi_{1\|2}$	n_{22} π_{22} $\pi_{2\|2}$	n_{23} π_{23} $\pi_{3\|2}$	n_{2+} π_{2+}
Total	n_{+1} π_{+1}	n_{+2} π_{+2}	n_{+3} π_{+3}	n_{++}

Table of Treatment by outcome

Treatment	outcome			
Frequency Percent Row Pct Col Pct	Fatal Attack	NonFatal Attack	No Attack	Total
Aspirin	n_{11} 5 π_{11} 0.02 $\pi_{1\|1}$ 0.05 21.74	n_{12} 99 π_{12} 0.45 $\pi_{2\|1}$ 0.90 36.67	n_{13} 10933 π_{13} 49.54 $\pi_{3\|1}$ 99.06 50.20	n_{1+} 11037 π_{1+} 50.01
Placebo	n_{21} 18 π_{21} 0.08 $\pi_{1\|2}$ 0.16 78.26	n_{22} 171 π_{22} 0.77 $\pi_{2\|2}$ 1.55 63.33	n_{23} 10845 π_{23} 49.14 $\pi_{3\|2}$ 98.29 49.80	n_{2+} 11034 π_{2+} 49.99
Total	n_{+1} 23 π_{+1} 0.10	n_{+2} 270 π_{+2} 1.22	n_{+3} 21778 π_{+3} 98.67	n_{++} 22071 100.00

Notation

- Joint distribution: π_{ij} $i=1,\ldots,I$; $j=1,\ldots,J$
  - the probability that (X, Y) occurs in the cell in row i and column j.

- marginal distributions: the row and column totals that result from summing the joint probabilities.
  - π_{i+} $i=1,\ldots,I$ for the row variable
  - π_{+j} $j=1,\ldots,J$ for the column variable

where $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$.

$\Sigma_i \pi_{i+} = \Sigma_j \pi_{+j} = \Sigma_i \Sigma_j \pi_{ij} = 1.0.$

Conditional probability

- one variable, say *Y, is a* response variable and the other *X is an explanatory variable.*
- for a fixed category of *X, Y has a probability* distribution.
- Given that a subject is classified in row *i of X,* $\pi_{j|i}$ denotes the probability of classification in column *j of Y, j=1, . . . , J.*
- The probabilities $\{\pi_{1|i},\ldots,\pi_{J|i}\}$ *form the conditional distribution* of *Y at category i of X* *where* $\sum_j \pi_{j|i} = 1.$

2.1.2 Sensitivity and Specificity

TABLE 2.2 Estimated Conditional Distributions for Breast Cancer Diagnoses

Breast Cancer	Diagnosis of Test		
	Positive	Negative	Total
Yes	0.82	0.18	1.0
No	0.01	0.99	1.0

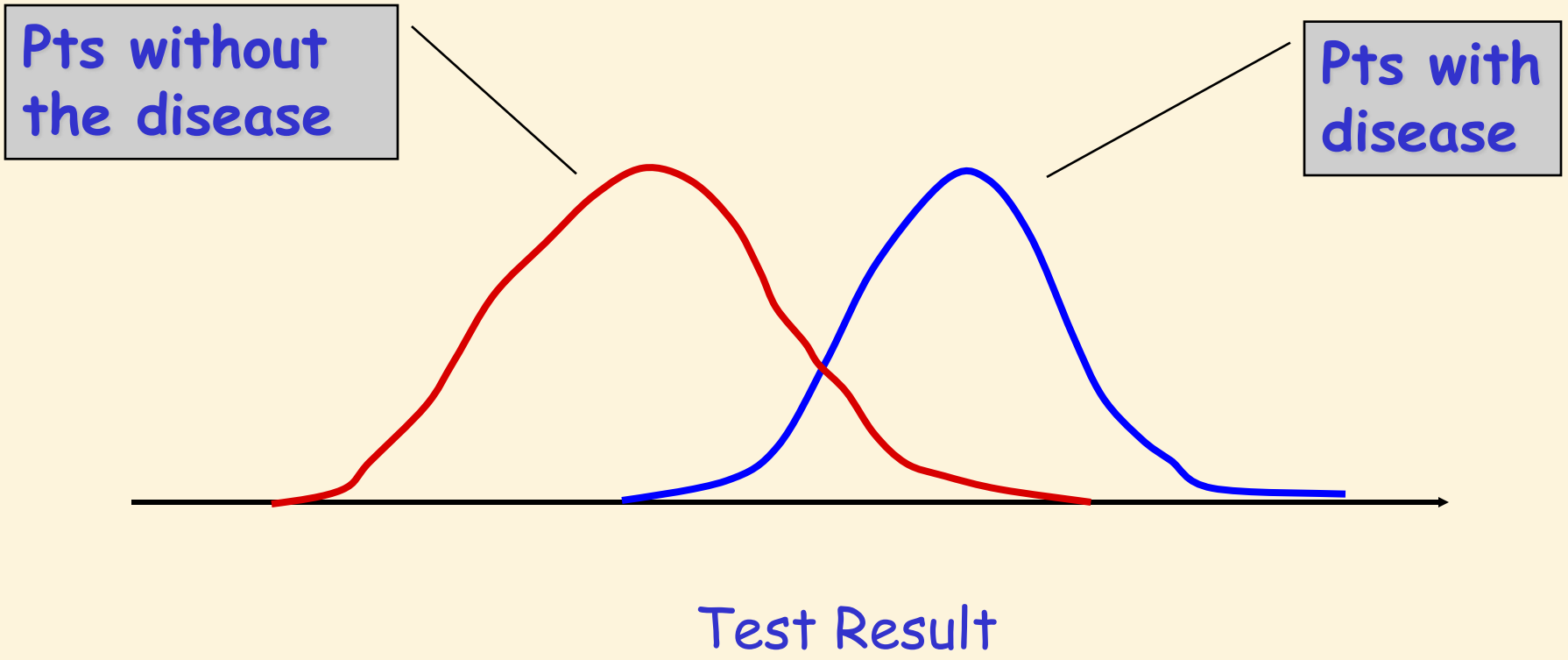
Source: Data from W. Lawrence et al., *J. Natl. Cancer Inst.* **90**: 1792–1800 (1998).

- Given that the subject has the disease, the conditional probability that the diagnostic test is positive is called the sensitivity;
- given that the subject does not have the disease, the conditional probability that the test is negative is called the specificity

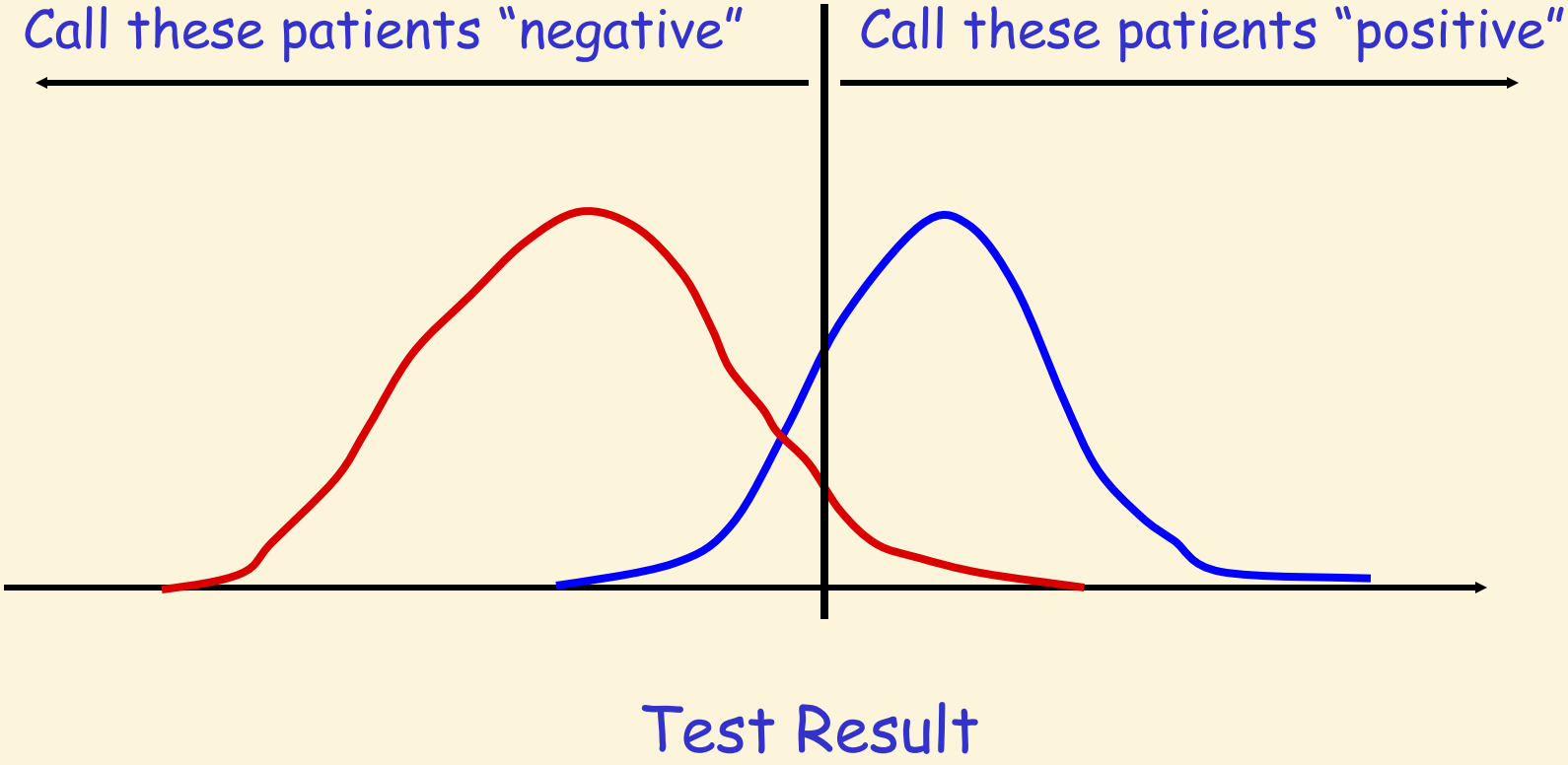
True disease state vs. Test result

Disease \ Test	rejected (+)	not rejected (-)
Disease (Yes)	☺ Power 1 - β; sensitivity $\pi_{1\|1}$	X Type II error (False -) β
No disease (No)	X Type I error (False +) α	☺ specificity $\pi_{2\|2}$

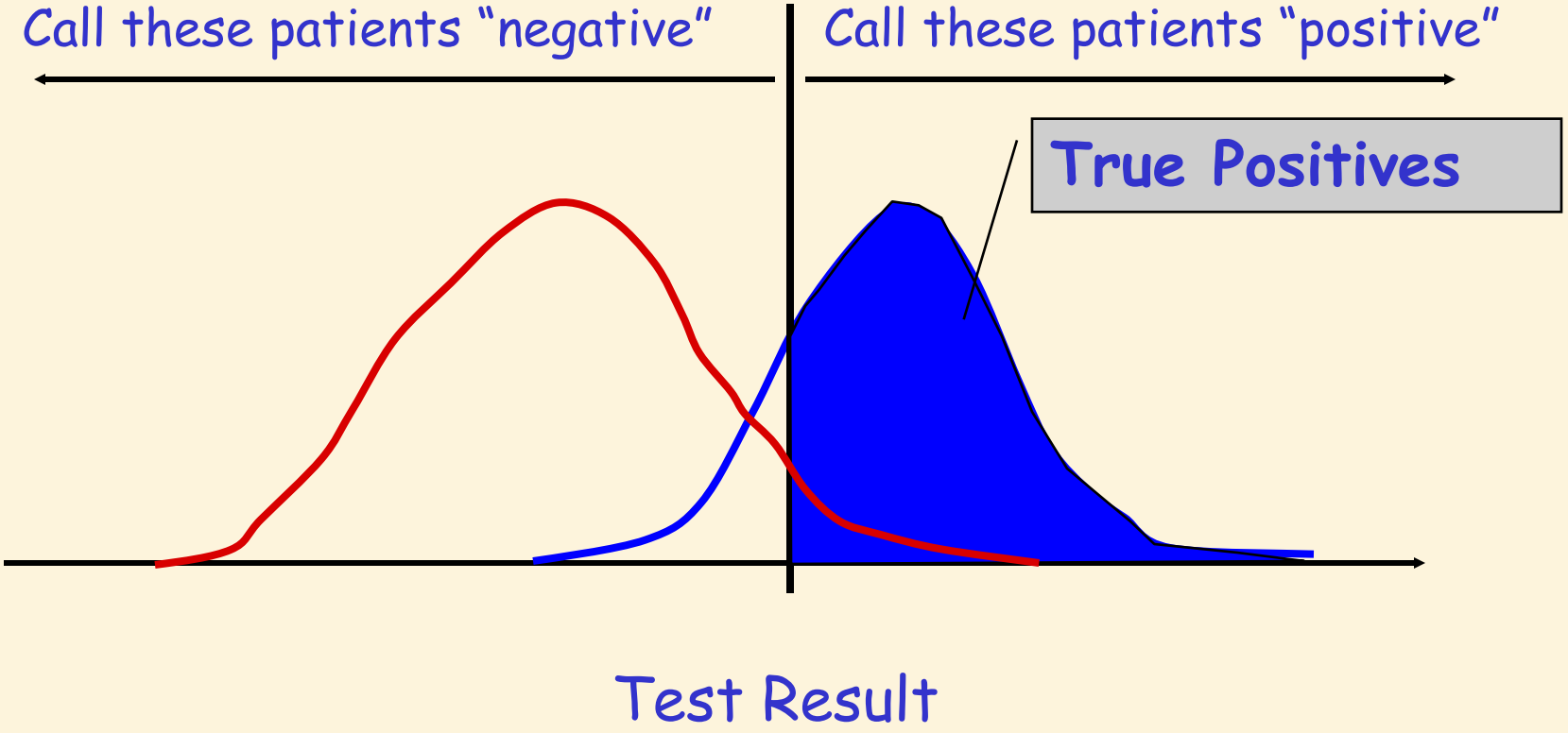
Specific Example



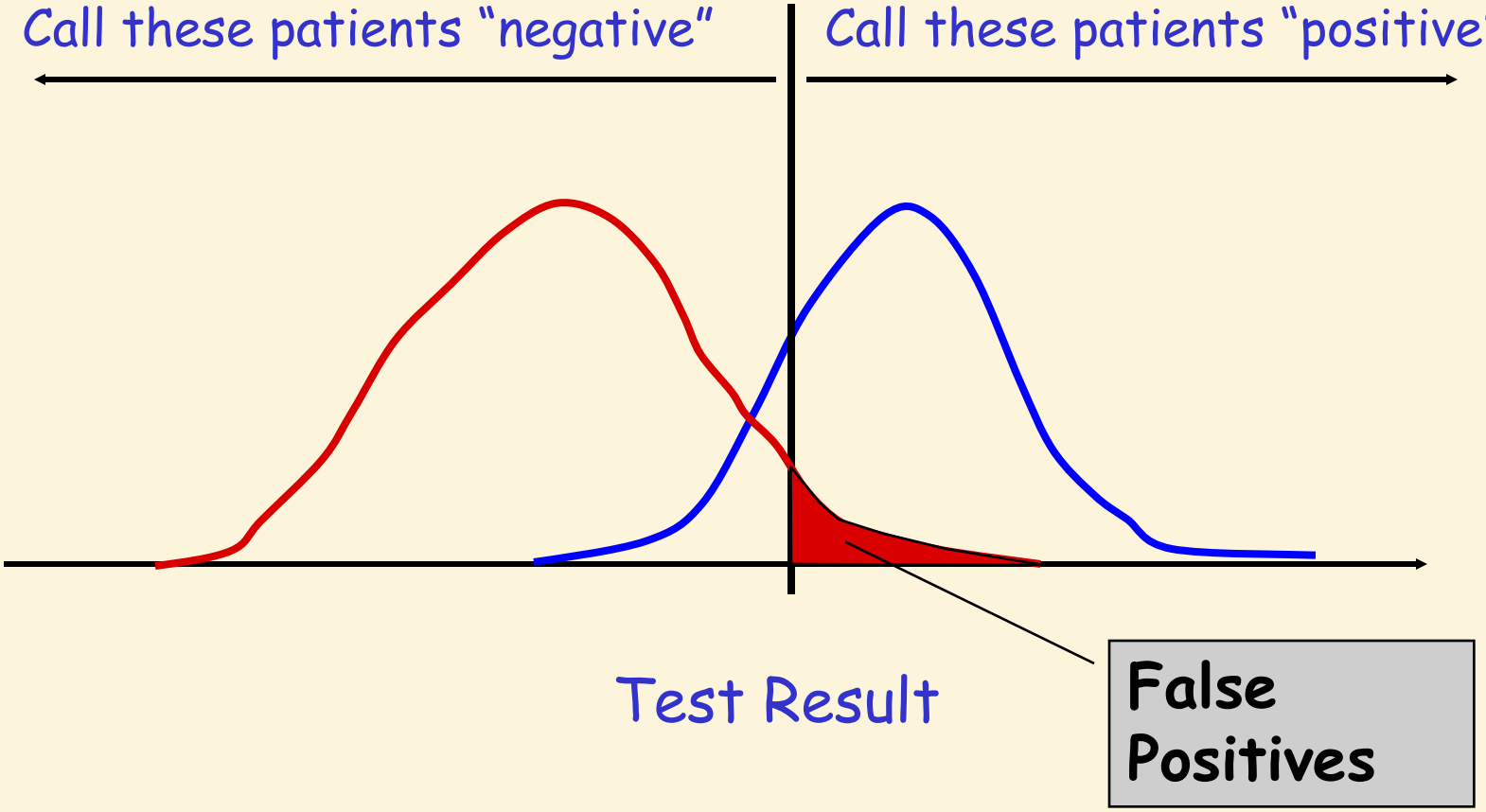
Threshold


Call these patients "negative"
Call these patients "positive"
Test Result

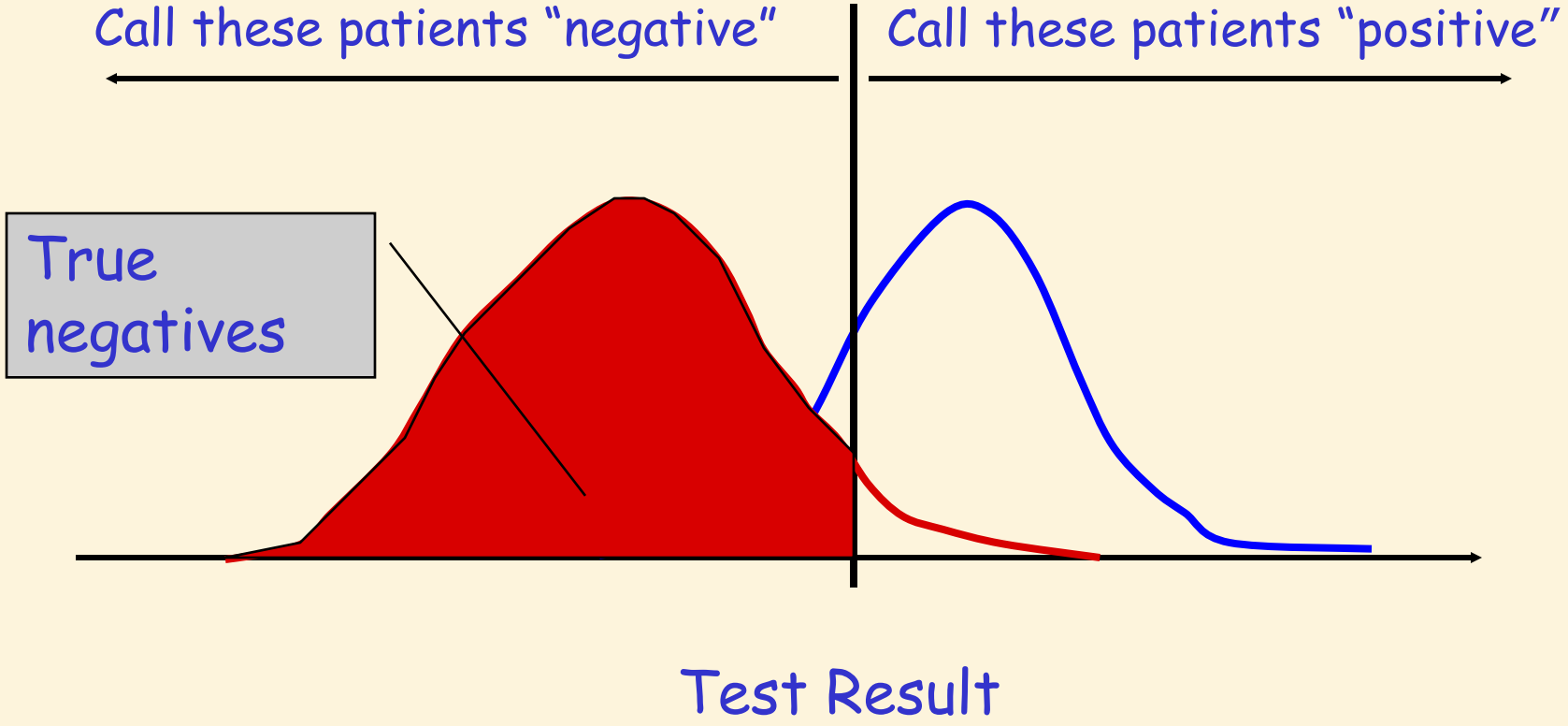
Some definitions ...



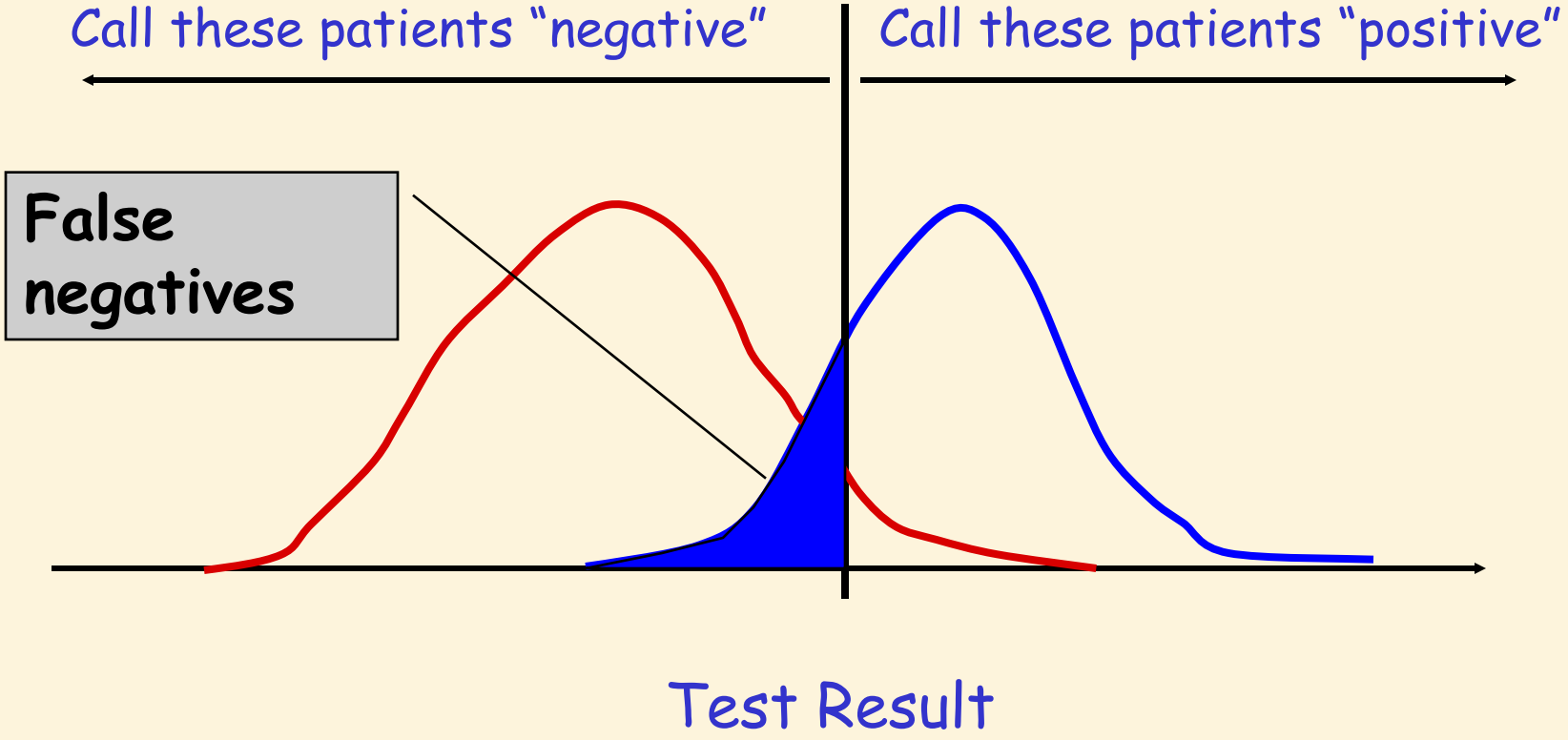
without the disease
with the disease



without the disease
with the disease

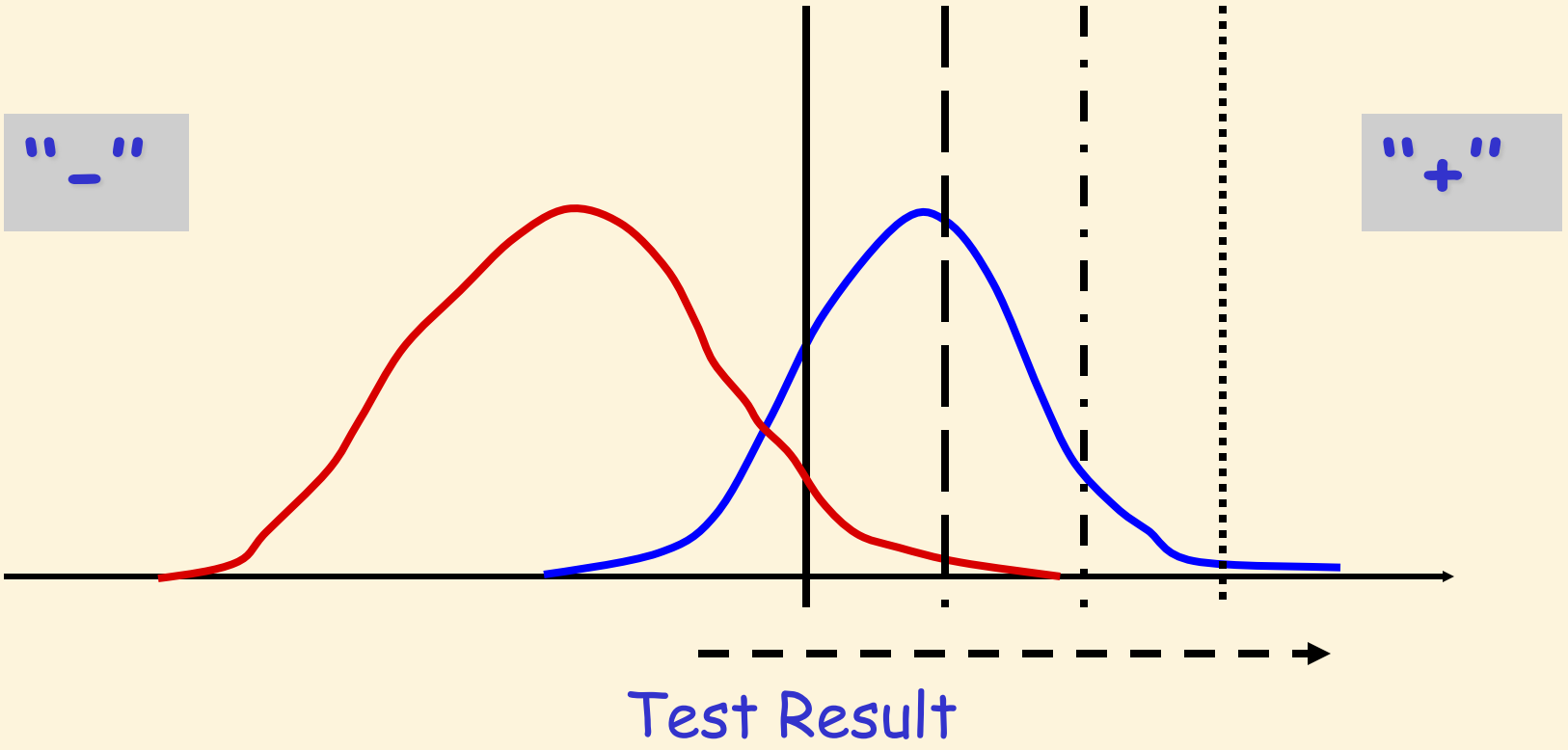


without the disease
with the disease



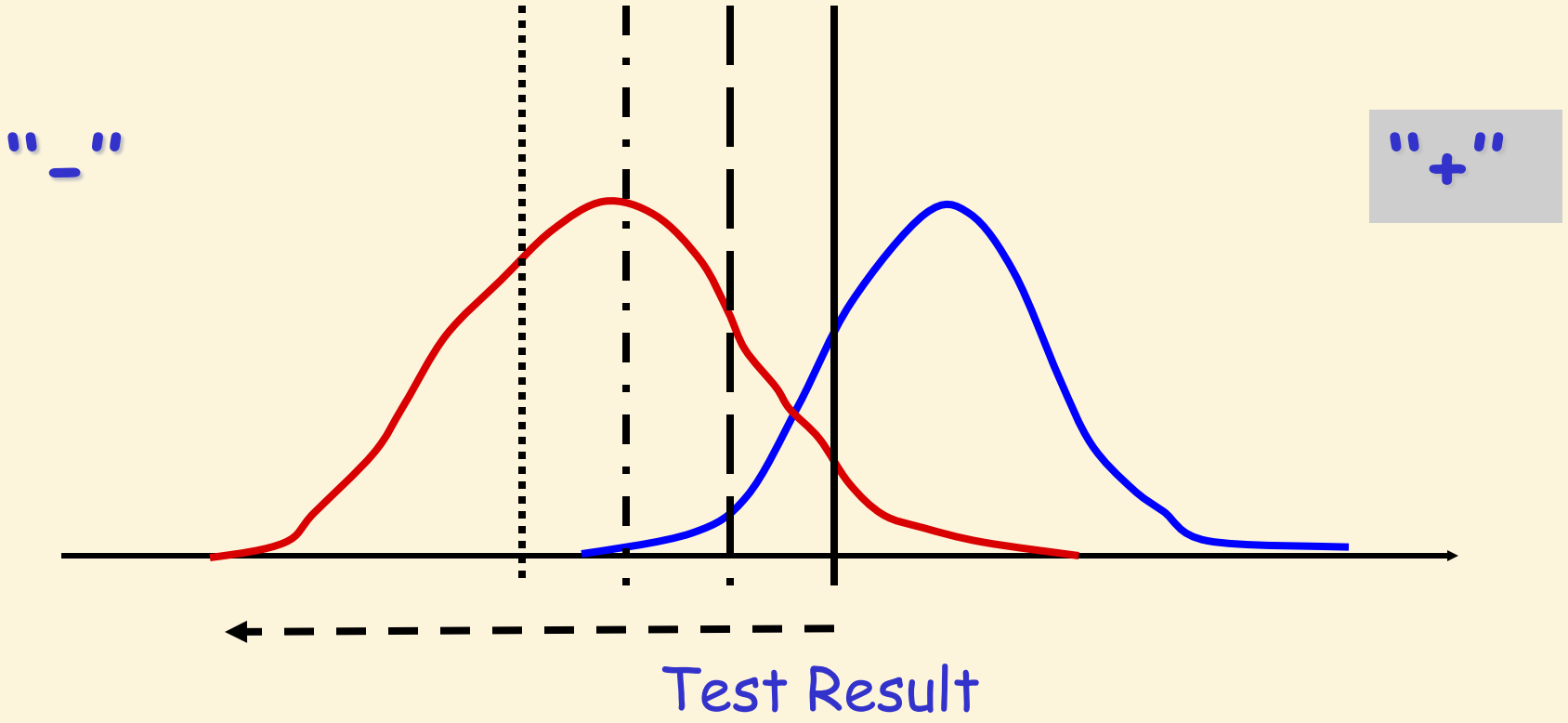
without the disease
with the disease

Moving the Threshold: right



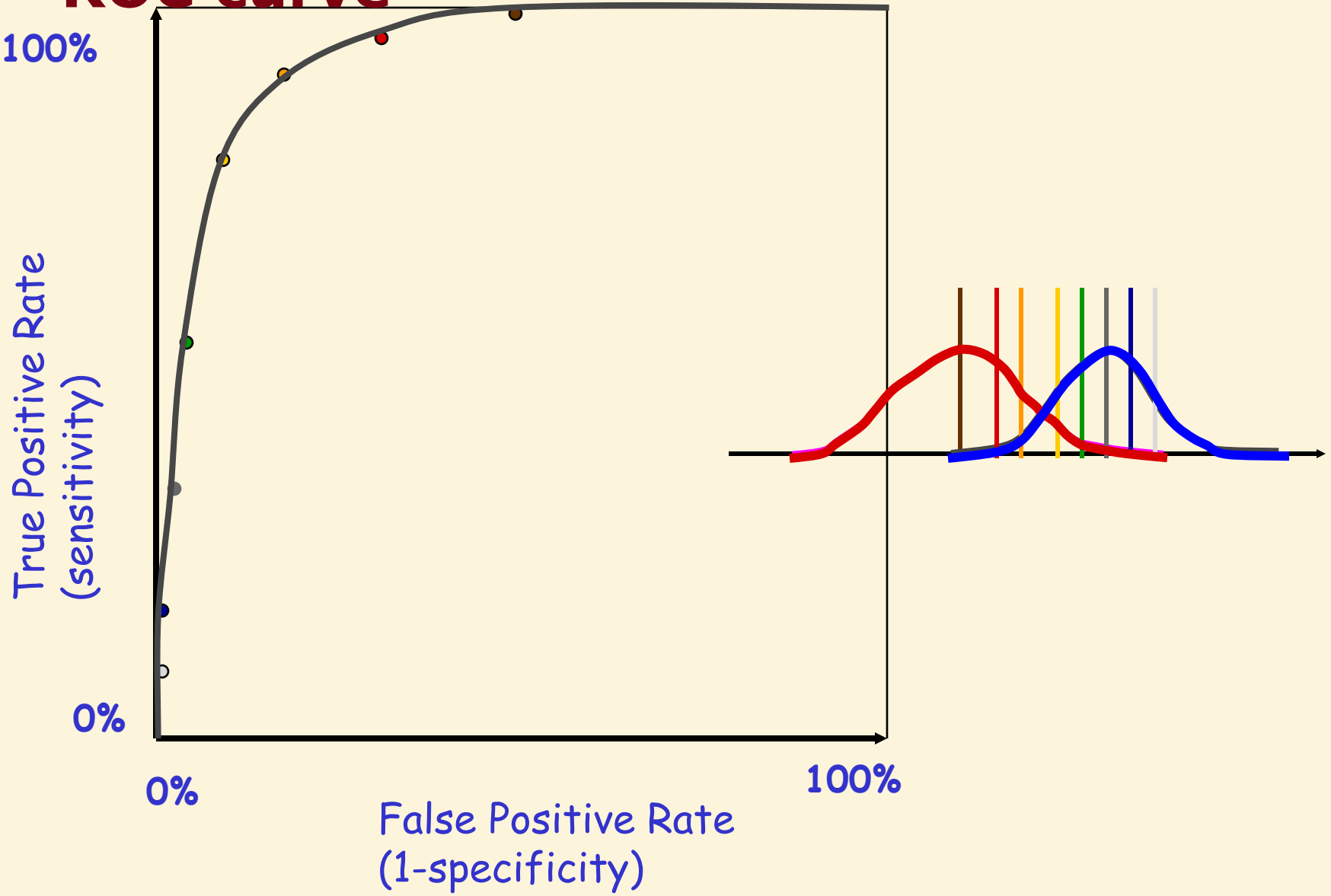
without the disease
with the disease

Moving the Threshold: left



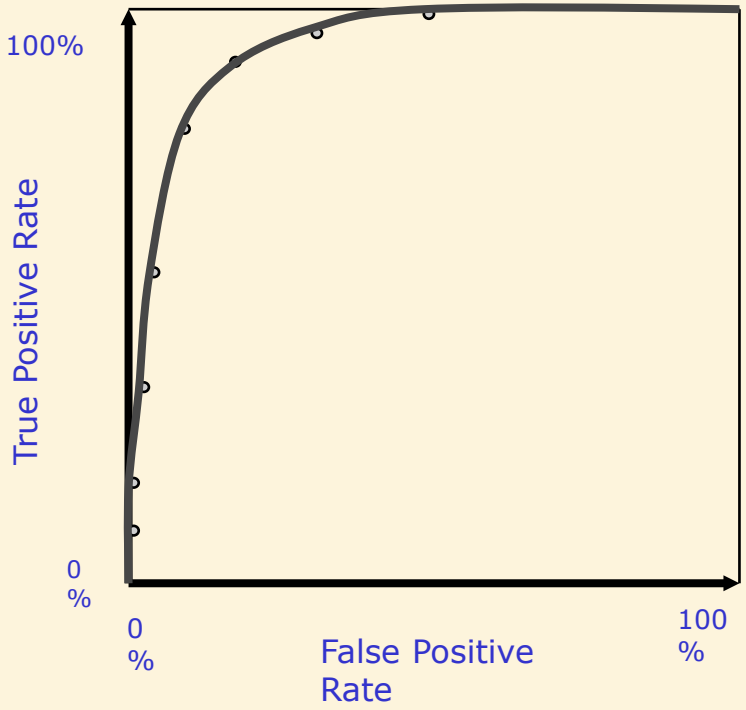
without the disease
with the disease

ROC curve


100%
True Positive Rate
(sensitivity)
0%
0%
100%
False Positive Rate
(1-specificity)

ROC curve comparison

A good test:



A poor test:

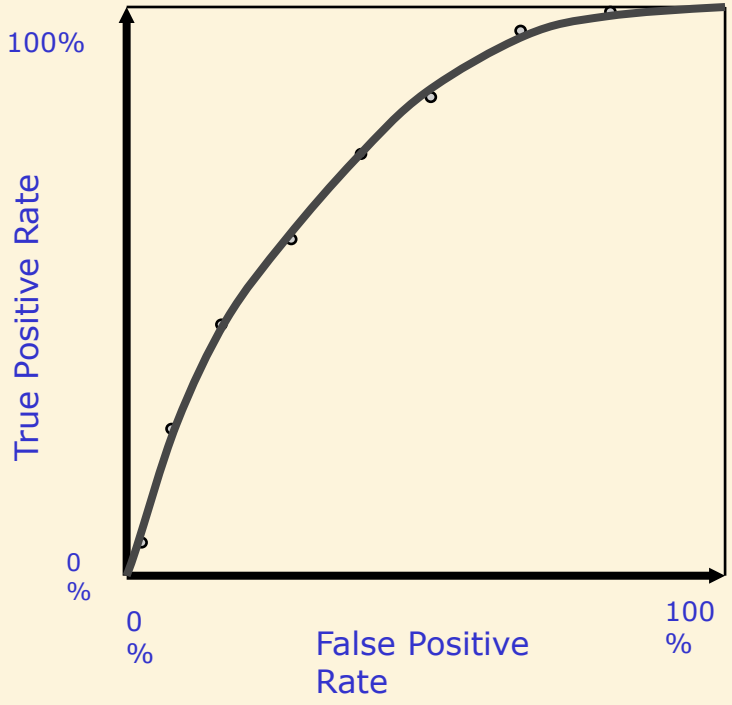


TABLE 2.2 Estimated Conditional Distributions for Breast Cancer Diagnoses

Breast Cancer	Diagnosis of Test		
	Positive	Negative	Total
Yes	0.82	0.18	1.0
No	0.01	0.99	1.0

Source: Data from W. Lawrence et al., *J. Natl. Cancer Inst.* **90**: 1792–1800 (1998).

- the estimated sensitivity of combined mammography and clinical examination is 0.82.
  - Of women with breast cancer, 82% are diagnosed correctly.
- the estimated specificity is 0.99.
  - Of women not having breast cancer, 99% were diagnosed correctly.

2.1.3 Independence of Categorical Variables

- In previous section, we defined *sensitivity* $\pi_{1|1}$ and *specificity* $\pi_{2|2}$. It is used to measure agreement of a test.
- Another important concern is the association between two categorical variables
- Or equivalently, the independence between two categorical variables
- Usually define it based on their joint distribution, the conditional distributions of Y given X, or of X given Y.

Independence

- Recall in probability:
  - two events A and B are independent if and only if P(AB)=P(A)*P(B)
  - Two random variables are independent if and only if f(x, y)=f(x)*f(y)
- Two categorical response variables are defined to be *independent if all* joint probabilities equal the product of their marginal probabilities,

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \text{for } i = 1, \ldots, I \quad \text{and} \quad j = 1, \ldots, J. \qquad (2.1)$$

- Thus, when *X and Y are independent,*

$$\pi_{j|i} = \pi_{ij}/\pi_{i+} = (\pi_{i+}\pi_{+j})/\pi_{i+} = \pi_{+j} \quad \text{for } i = 1, \ldots, I.$$

Independence

- Each conditional distribution of *Y is identical to the marginal distribution of Y.*
- Or for any j=1,…,I, $\pi_{j|1} = \cdots = \pi_{j|I}$

- Independence is then often referred to as *homogeneity of the conditional distributions.*

Notations

Population parameters	Sample
π	p or $\hat{\pi}$
π_{ij}	$p_{ij} = n_{ij}/n.$
$\pi_{j\|i}$	$p_{j\|i} = p_{ij}/p_{i+} = n_{ij}/n_{i+}$

where $n = \Sigma_i \Sigma_j n_{ij}$

$n_{i+} = np_{i+} = \Sigma_j n_{ij}.$

2.1.4 Poisson, Binomial, and Multinomial Sampling

- The probability distributions introduced in Section 1.2 extend to cell counts in contingency tables.
  - Poisson
  - Binomial
  - Multinomial

Poisson

- a Poisson sampling model treats cell counts $\{Y_{ij}\}$ *as independent Poisson random variables with parameters* $\{\mu_{ij}\}$.
- The joint probability mass function for potential outcomes $\{n_{ij}\}$ *is then the* product of the Poisson probabilities $P(Y_{ij} = n_{ij})$ *for the IJ cells, or*

$$\prod_i \prod_j \exp(-\mu_{ij})\,\mu_{ij}^{n_{ij}}/n_{ij}!\,.$$

Multinomial Sampling

- When the total sample size *n is fixed but the row and column totals are* not, a *multinomial sampling model applies.*
- *The IJ cells are the possible* outcomes.
- The probability mass function of the cell counts has the multinomial form

$$[n!/(n_{11}! \cdots n_{IJ}!)] \prod_i \prod_j \pi_{ij}^{n_{ij}}.$$

product multinomial sampling

- Observations on a response *Y occur separately at each setting of an* explanatory variable *X.*
- *This case normally treats row totals as fixed, and for* simplicity, we use the notation $n_i = n_{i+}$.
- *Suppose that the n_i observations on Y at setting i of X are independent, each with probability distribution*

$$\{\pi_{1|i}, \ldots, \pi_{J|i}\}$$

- *The counts* $\{n_{ij},\ j = 1, \ldots, J\}$ *satisfying* $\sum_j n_{ij} = n_i$ *then have* the multinomial form

$$\frac{n_i!}{\prod_j n_{ij}!} \prod_j \pi_{j|i}^{n_{ij}}.$$

- When samples at different settings of *X are independent, the joint probability* function for the entire data set is the product of the multinomial functions

Hypergeometric distribution

- Sometimes both row and column margins are naturally fixed.
- The appropriate sampling distribution is then the *hypergeometric.*
- *It* is less common.

2.1.5 Seat Belt Example

- Researchers in the Massachusetts Highway Department plan to study
  - the relationship between seat-belt use (yes, no) and outcome of an automobile crash (fatality, nonfatality) for drivers involved in accidents on the Massachusetts Turnpike.
- Data is summarized as

TABLE 2.4 Seat-Belt Use and Results of Automobile Crashes

	Result of Crash	
Seat-Belt Use	Fatality	Nonfatality
Yes		
No		

Seat Belt Example

- Design 1: They plan to catalog all accidents on the turnpike for the next year, classifying each according to these variables.
  - The total sample size is then a random variable
  - Treat the numbers of observations at the four combinations as independent Poisson $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$

- Design 2: researchers randomly sample 200 police records of crashes on the turnpike in the past year and classify each according to seat-belt use and outcome of crash.
  - n=200 is fixed
  - Multinomial (n, $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$)

Seat Belt Example

- Design 3: The researchers might instead randomly sample 100 records of accidents with a fatality and randomly sample 100 records of accidents with no fatality.
  - This approach fixes the column totals in Table 2.4 at 100.
  - They might then regard each column of Table 2.4 as an independent binomial sample. (retrospective)
- Design 4: (traditional experimental design)
  - takes 200 subjects and randomly assigns 100 of them to wear seat belts; the 200 then all are forced to have an accident. (prospective)
  - The recorded results would then be independent binomial samples in each row, with fixed row totals of 100 each. May be unethical for humans.

2.1.6 Study Design

- **Observational studies**
  - Cohort study
    - Prospective cohort
    - Retrospective cohort
    - Time series study
  - Case-control study
  - Cross-sectional study
- **Treatment studies (experimental studies, prospective)**
  - Randomized controlled trial
    - Double-blind randomized trial
    - Single-blind randomized trial
    - Non-blind trial
  - Nonrandomized trial (quasi-experiment)

Cohort study

- A cohort study or panel study is a form of longitudinal study used in medicine and social science. It is one type of study design and should be compared with a cross-sectional study.
- A cohort is a group of people who share a common characteristic or experience within a defined period (e.g., are born, leave school, lose their job, are exposed to a drug or a vaccine, etc.).

prospective cohort study

- A **prospective cohort study** is a research effort that
  - follows over time groups of individuals
  - who are similar in some respects (e.g., all are working adults)
  - but differ on certain other characteristics (e.g., some smoke and others do not)
  - and compares them for a particular outcome (e.g., lung cancer)
- It can be more expensive than a case–control study.

retrospective cohort study

- A **retrospective cohort study**, also called a **historic cohort study**, is a medical research study in which
  - the medical records of groups of individuals
  - who are alike in many ways
  - but differ by a certain characteristic (for example, female nurses who smoke and those who do not smoke)
  - are compared for a particular outcome (such as lung cancer).
- As is obvious, Retrospective Cohort has the benefits of being cheaper and less time consuming.
- The resources are mainly directed at collection of data only.

Case-control study

- **Case-control** is a type of epidemiological study design.
  - Case-control studies are used to identify factors that may contribute to a medical condition
  - by comparing subjects who have that condition (the 'cases') with patients who do not have the condition (the 'controls')
  - but are otherwise similar.
- Case-control studies are a relatively inexpensive and frequently-used type of epidemiological study that can be carried out by small teams

A case control study

- In 20 hospitals in London, England, patients admitted with lung cancer in the preceding year were queried about their smoking behavior.
- For each of the 709 patients admitted, researchers studied the smoking behavior of a noncancer patient at the same hospital of the same gender and within the same 5-year grouping on age.

TABLE 2.5 Cross-Classification of Smoking by Lung Cancer

	Lung Cancer	
Smoker	Cases	Controls
Yes	688	650
No	21	59
Total	709	709

Source: Based on data reported in Table IV, R. Doll and A. B. Hill, *British Med. J.*, Sept. 30, 1950, pp. 739–748.

lung cancer with smoking behavior

- distribution of lung cancer is fixed by the sampling design, and the outcome measured is whether the subject ever was a smoker.
- It is a retrospective design to "look into the past"
- Often, the two samples are matched, as in this study.
- Sometimes the samples of cases and controls are independent rather than matched.
- For those in Table 2.5 with lung cancer, the proportion who were smokers was 688/709=0.970, while it was 650/709=0.917 for the controls.
- we cannot estimate the probability of lung cancer at each category of smoking behavior, without knowing the proportion of the population having lung cancer.

Cross-sectional study

- **Cross-sectional** form a class of research methods that
  - involve observation of some subset of a population of items all at the same time,
  - in which, groups can be compared at different ages with respect of independent variables, such as IQ and memory.
- The fundamental difference between cross-sectional and longitudinal studies is that cross-sectional studies take place at a single point in time and that a longitudinal study involves a series of measurements taken over a period of time. Both are a type of observational study.

Clinical trial

- A **randomized controlled trial (RCT)** is a type of scientific experiment most commonly used in testing the efficacy or effectiveness of healthcare services (such as medicine or nursing) or health technologies (such as pharmaceuticals, medical devices or surgery).
- RCTs are also employed in other research areas, such as judicial, educational, and social research.
- As their name suggests, RCTs involve the random allocation of different interventions (treatments or conditions) to subjects.
- As long as numbers of subjects are sufficient, this ensures that both known and unknown confounding factors are evenly distributed between treatment groups.

Clinical trial

- Open-label trial: the researcher knows the full details of the treatment, and so does the patient.
- Blind trials
  - **Single-blind trial:** the researcher knows the details of the treatment but the patient does not
  - **Double-blind trial:** the researcher and patient "does not" know the details of the treatment
  - **Triple-blind trial:** it may mean that the patient, researcher and statistician are blinded.
- Double-blind trials are preferred, as they tend to give the most accurate results.

Distribution assumption

- Prospective studies usually condition on the totals $\{n_i = \Sigma_j \ n_{ij}\}$ for categories of X and regard each row of J counts as an independent multinomial sample on Y.
- Retrospective studies usually treat the totals $\{n_{+j}\}$ for Y as fixed and regard each column of I counts as a multinomial sample on X.
- In cross-sectional studies, the total sample size is fixed but not the row or column totals, and the IJ cell counts are a multinomial sample.