

Biostatistics

Week #4 (3/24/2020)



Chapter 6 Probability and Diagnostic Tests



Outline

- 6.1 Operations on Events and Probability
- 6.2 Conditional Probability
- 6.3 Bayes' Theorem
- 6.4 Diagnostic Tests
 - Sensitivity and Specificity
 - Application of Bayes' Theorem
 - ROC Curves
 - Prevalence evaluation
- ~~6.5 The Relative Risk (RR) and the Odds Ratio (OR)~~

Preface

- Previously we learned various techniques in “**descriptive**” statistics.
- In addition to describing these “observations”, we are also interested in investigating how the information contained in the sample can be used to “**infer**” the characteristics of the population from which it was drawn.
- **Probability** theory lays the foundation for such discussion.

6.1 Operations on Events and Probability

- An event is the basic element to which probability can be applied.
 - It is the result of an observation or experiment. For example, the event that “a 30-year-old woman lives to see her 70th birthday”, or the event that “the same woman is diagnosed with cervical cancer before she reaches the age of 40”.

Probability Definition

- If an experiment is repeated n times under **essentially identical conditions**, and if event A occurs m times, then the probability of A is defined as :

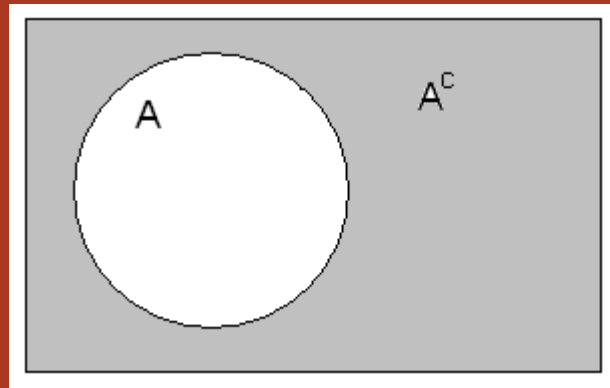
$$P(A) = \frac{m}{n}$$

- It is the relative frequency of occurrence – or the proportion of times the event occurs.
- **The numerical value of a probability lies between 0 and 1**
- An event never occurs is called a **null event**.

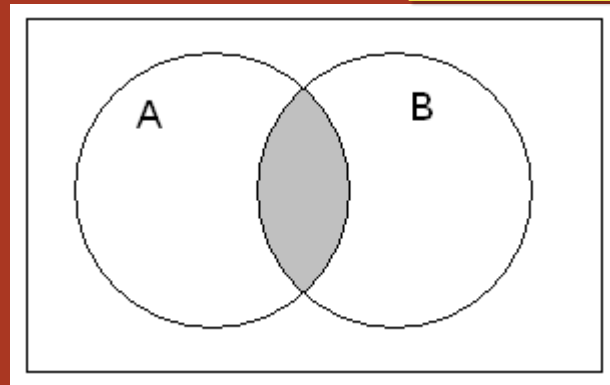
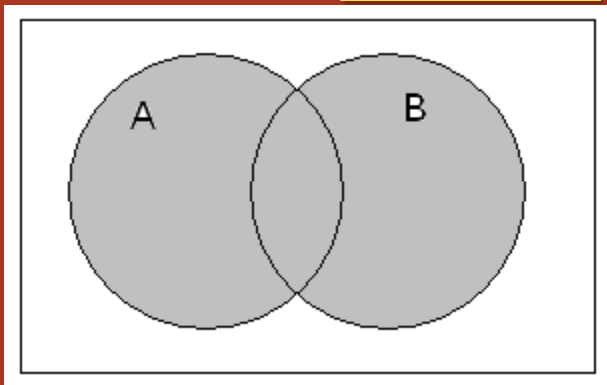
Probability Operations

- The probability of the **complementary** event A^c :

$$P(A^c) = 1 - \frac{m}{n}$$



- $P(A \cup B)$: **union**
- $P(A \cap B)$: **intersection**



6.2 Conditional Probability

- Given two events A and B , sometimes we also want to know whether **the prior occurrence of A causes** the probability of B to change.
- For example, we may find the probability that a person will live to the age of 65 (event A).
- We may also want to know the probability a person will survive for the next 5 years (event B) **once he or she has already reached the age of 65** (event A).

- The probability of event B occurs given that we know the outcome of event A :

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

given that $P(A) \neq 0$

- Similarly, we may have

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

given that $P(B) \neq 0$

(and is read "the conditional probability of B , given A " or "the probability of B under the condition A ".)

- If A and B are *independent* events (e.g., roll a dice to get 1, and do it again to get 1), then

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

- In this case, it comes natural that

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

as well as that $P(A | B) = P(A)$

The prior occurrence of A does NOT affect the occurrence of B (and vice versa).

6.3 Bayes' Theorem (wiki)

- Bayes' theorem relates the conditional probabilities of events A and B , where B has a non-vanishing probability:

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

- Similarly, when A has a non-vanishing probability:

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

Cont'd

- Intuitively, Bayes' theorem in this form describes the way in which one's beliefs about observing 'A' are updated by having observed 'B'.

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

Deriving Bayes' theorem

Earlier we know that

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

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

As a result:

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

This leads to:

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

or

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

Bayes' Theorem (in the text)

- If A_1, A_2, \dots, A_n are n mutually **exclusive** and **exhaustive**¹ events such that

$$P(A_1 \cup A_2 \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

- Example – rolling a dice to have 1, 2, 3, 4, 5 or 6.

¹The probabilities for all these mutually exclusive events **sum to 1**, that is, there are no other possible outcomes.

Bayes' Theorem – cont'd

- Bayes' theorem states that

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + \dots + P(A_n)P(B | A_n)}$$

for each i , $1 \leq i \leq n$ $A_1, A_2, \dots, \text{and } A_n$ are n mutually **exclusive** and **exhaustive** events

- For example:

$$P(A_1 | B) = \frac{P(A_1)P(B | A_1)}{P(A_1)P(B | A_1) + P(A_2)P(B | A_2)}$$

Bayes' Theorem – cont'd

- Bayes' theorem states that

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + \dots + P(A_n)P(B | A_n)}$$

for each i , $1 \leq i \leq n$

From previous slide:

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

Solution strategy

- Clearly define events A and B (or events $A_1, A_2 \dots$ and B) and their probabilities (with no prior conditions)
- Get the correct Bayes' theorem formula
- Obtain known conditional probabilities
- Solve for unknown conditional probability (or probabilities) of interest

Example #1

- Given the following conditions:
 - A school has 60% boys and 40% girls.
 - Girls wear trousers or skirts in equal number (50/50), while all boys wear trousers.
 - An observer sees a (random) student wearing trousers from a distance.
- What is the probability this student is a girl?

Assume we have 100 students, 40 are girls and 60 are boys (thus satisfying the first condition stated in previous slide).

| | Girls | Boys | Total |
|----------|-------|------|-------|
| Trousers | | | |
| Skirts | | | |
| Total | 40 | 60 | 100 |

(Setting up a table like this is not needed, but sometimes makes it easier for going along with this query.)


Based on the second condition that “Girls wear trousers or skirts in equal number (50/50), while all boys wear trousers”, we may complete the table as:

| | Girls | Boys | Total |
|----------|-------|------|-------|
| Trousers | 20 | 60 | 80 |
| Skirts | 20 | 0 | 20 |
| Total | 40 | 60 | 100 |

| | Girls | Boys | Total |
|----------|-------|------|-------|
| Trousers | 20 | 60 | 80 |
| Skirts | 20 | 0 | 20 |
| Total | 40 | 60 | 100 |

We want to know the probability for a trouser-wearing student to be a girl.

- Without using any formula, you can easily get the answer from this table.
- There are a total of 80 students wearing trousers. Anyone of them could be seen by you from a distance.
- There are 20 girls wearing trousers.
- As a result, the probability would be $20/80=0.25$.

- 
- When the number of events grows large, the setting up of such table becomes tedious.
 - Using Bayes' theorem is not only straightforward, it is easy to be programmed in computer.

| | Girls | Boys | Total |
|----------|-------|------|-------|
| Trousers | 20 | 60 | 80 |
| Skirts | 20 | 0 | 20 |
| Total | 40 | 60 | 100 |

We want to know the probability for a trouser-wearing student to be a girl.

- Let event A be the student observed is a **girl**, regardless other information. Thus $P(A)=0.4$.
- Let event B be the student wearing **trousers**, regardless other information. Thus $P(B)=0.8$.
- The probability of a trouser-wearing student to be a girl is $P(A|B)$. According to Bayes' theorem:

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

We already know $P(A)$ and $P(B)$.
What is $P(B|A)$?

| | Girls | Boys | Total |
|----------|-------|------|-------|
| Trousers | 20 | 60 | 80 |
| Skirts | 20 | 0 | 20 |
| Total | 40 | 60 | 100 |

We want to know the probability for a trouser-wearing student to be a girl.

- We know $P(B|A)=0.5$ from the above table. (Girls wearing trousers are 50% among girls.)

$$P(A | B) = \frac{P(A)P(B | A)}{P(B)} = 0.25$$

0.4 0.5
0.8

Solution strategy (refresh)

- Clearly define events A and B (or events $A_1, A_2 \dots$ and B) and their probabilities (*with no prior conditions*)
- Get the correct Bayes' theorem formula
- Obtain known conditional probabilities
- Solve for unknown conditional probability (or probabilities)

- Let event A be the student observed is a girl, **regardless** other information.
- Let event B be the student wearing trousers, **regardless** other information.

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

$$P(A) = 0.4. \quad P(B) = 0.8.$$
$$P(B|A) = 0.5$$

$$\begin{aligned} P(A|B) &= P(B|A)P(A) / P(B) \\ &= 0.5 \times 0.4 / 0.8 \\ &= 0.25. \end{aligned}$$

Diagnosis (wiki)

- **Diagnosis** is the identification of the nature and cause of a certain phenomenon.
- Diagnosis is used in many different disciplines with variations in the use of logic, analytics, and experience to determine "cause and effect".
- In systems engineering and computer science, it is typically used to determine the causes of symptoms, mitigations*, and solutions.

*the effort to reduce loss of life and property by lessening the impact of disasters. (mitigate: 減輕)

6.4 Diagnostic Tests

- A **diagnostic test** is any kind of medical test performed to aid in the diagnosis (診斷) or detection of disease. For example:
 - to diagnose diseases
 - to measure the progress or recovery from disease
 - to confirm that a person is free from disease



Cont'd

- The result of a test may be **positive (陽性)** or **negative (陰性)**: this has nothing to do with **a bad prognosis***, but rather means that the test worked or not.
- For example, a negative screening test for breast cancer means that no sign of breast cancer could be found (which is in fact very positive for the patient).

* the doctor's ***prediction*** of how a patient will progress, and whether there is a chance of recovery.

Cont'd

- Bayes' theorem is often used to derive the probability for an individual to develop into a particular disease who has not yet exhibited any clinical symptoms.

TP, FP, TN and FN

- The test outcome can be positive (sick) or negative (healthy), while the actual health status of the persons may be different. In that setting:
 - True positive (TP or 陽性): Sick people correctly identified as sick
 - False positive (FP or 偽陽性): Healthy people wrongfully identified as sick
 - True negative (TN or 陰性): Healthy people correctly identified as healthy
 - False negative (FN or 偽陰性): Sick people wrongfully identified as healthy

Sensitivity & Specificity

- Sensitivity (靈敏度) and specificity (特異性) are statistical measures of the performance of a binary classification test.
- The sensitivity measures the proportion of actual positives which are correctly identified (e.g. the percentage of sick people who are identified as having the condition)

$$sensitivity = \frac{TP}{TP + FN}$$

Cont'd

- The specificity, on the other hand, measures the proportion of actual negatives which are correctly identified (e.g. the percentage of well people who are identified as not having the condition)

$$specificity = \frac{TN}{TN + FP}$$

PPV and NPV

- Positive Predictive Value (陽性預測值):
the proportions of positive results that
are true positive results.
- As a result $PPV = TP / (TP + FP)$
- Negative Predictive Value (陰性預測值):
the proportions of negative that are true
negative results.
- As a result $NPV = TN / (TN + FN)$

Example #2

- A Fecal occult blood (FOB; 糞便潛血) screen test is used in 203 people to look for bowel cancer (大腸癌).
 - 20 were tested positive
 - 183 were tested negative
- Among those 20 tested positive, only 2 were confirmed with cancer (by further medical examination).
- Among those 183 negative cases, 1 was unfortunately a true cancer case.

Cont'd

- Determine TP, TN, FP and FN of this test.
- Determine the sensitivity and specificity of this test.
- Determine the positive predictive value (PPV) and negative predictive value (NPV) of this test.

- Among those 20 tested positive, only 2 were confirmed with cancer (by further medical examination).
- Among those 183 negative cases, 1 was unfortunately a true cancer case.

| | | Patients with <u>bowel cancer</u> (as confirmed on <u>endoscopy</u>) | |
|---------------------------------|------------------------|--|--|
| F O B t e s t | <i>Posi- tive</i> | TP = <input data-bbox="1031 714 1155 865" type="text" value="?"/> | FP = <input data-bbox="1408 714 1532 865" type="text" value="?"/> |
| | <i>Nega- -tive</i> | FN = <input data-bbox="1031 986 1155 1138" type="text" value="?"/> | TN = <input data-bbox="1396 986 1539 1138" type="text" value="?"/> |

- Among those 20 tested positive, only 2 were confirmed with cancer (by further medical examination).
- Among those 183 negative cases, 1 was unfortunately a true cancer case.

| | | Patients with <u>bowel cancer</u> (as confirmed on <u>endoscopy</u>) | |
|---------------------------------|-----------------------|--|----------|
| F O B t e s t | <i>Posi- tive</i> | TP = 2 | FP = 18 |
| | <i>Nega- tive</i> | FN = 1 | TN = 182 |

| | | Patients with <u>bowel cancer</u> (as confirmed on <u>endoscopy</u>) | | |
|--|------------------------------|--|------------------------------------|---|
| F O B t e s t | <i>Posi- tive</i> | TP = 2 | FP = 18 | PPV (Positive Predictive Value) <div>?</div> |
| | <i>Nega- tive</i> | FN = 1 | TN = 182 | NPV (Negative Predictive Value) <div>?</div> |
| | | Sensitivity <div>?</div> | Specificity <div>?</div> | |

| | | Patients with <u>bowel cancer</u> (as confirmed on <u>endoscopy</u>) | | |
|--|------------------------------|---|---|--|
| F O B t e s t | <i>Posi- tive</i> | TP = 2 | FP = 18 | PPV (Positive Predictive Value) = TP / (TP + FP) = 2 / (2 + 18) = 2 / 20 = 10% |
| | <i>Nega- tive</i> | FN = 1 | TN = 182 | NPV (Negative Predictive Value) = TN / (TN + FN) = 182 / (1 + 182) = 182 / 183 = 99.5% |
| | | Sensitivity = TP / (TP + FN) = 2 / (2 + 1) = 2 / 3 = 66.67% | Specificity = TN / (FP + TN) = 182 / (18 + 182) = 182 / 200 = 91% | |

Discussion - Sensitivity

- Sensitivity = $2 / (2+1) = 66.7\%$
 - Out of 100 cancer patients, this test is capable of detecting 67 patients.
 - In other words, 33 patients might be wrongfully diagnosed as healthy.
 - *If you were tested positive (i.e., you have cancer), are you convinced?*
 - *Are you worried if you were tested negative?*

Discussion – Specificity

- Specificity = $182 / (182 + 18) = 91.0\%$
 - We are 91% sure that we are healthy if the test result is negative.
 - Out of 100 healthy individuals, there are 9 who might be wrongfully diagnosed as cancer.
 - *If you were tested negative (i.e., you are free from cancer), are you convinced?*
 - *Are you worried if you were the other 9%?*

Which one is better?

- Wrongfully diagnosed as cancer?
(9% chance)
 - You will probably take further medical exams, with strong belief that these tests will prove that you are healthy.
- Wrongfully diagnosed as healthy?
(33% chance)
 - I am not sure that I am healthy?
 - No further examinations to follow?

Contradiction?

- Should I be happy if I were tested having cancer, or not having cancer?
- In general an FOB test is used as a screening (篩選) method, with the purpose of revealing more potential cancer patients.
- Although its PPV is only 10% (only 2 out of 20 positive cases are cancer patients), the identification of them will help in promoting their survivals as well as reducing long-term medical expenses.

Cont'd

- On the other hand, it may not be a good diagnostic method for an individual (as we have seen from previous discussion).
- Many diseases with **low prevalence** follow this pattern (the FOB test).
- Increase sample size will help (see next example).

6.4.2 Application of Bayes' Theorem – Example #3

- Pap smear (子宮頸抹片檢查) is a test among women to check if they have cervical cancer (子宮頸癌).
- We have seen from previous example, that among 20 individuals who were tested positive from FOBT, only 2 (10%) were actually bowel cancer patients.
- What is the probability that a woman with a Pap smear positive, and actually does have the cancer?

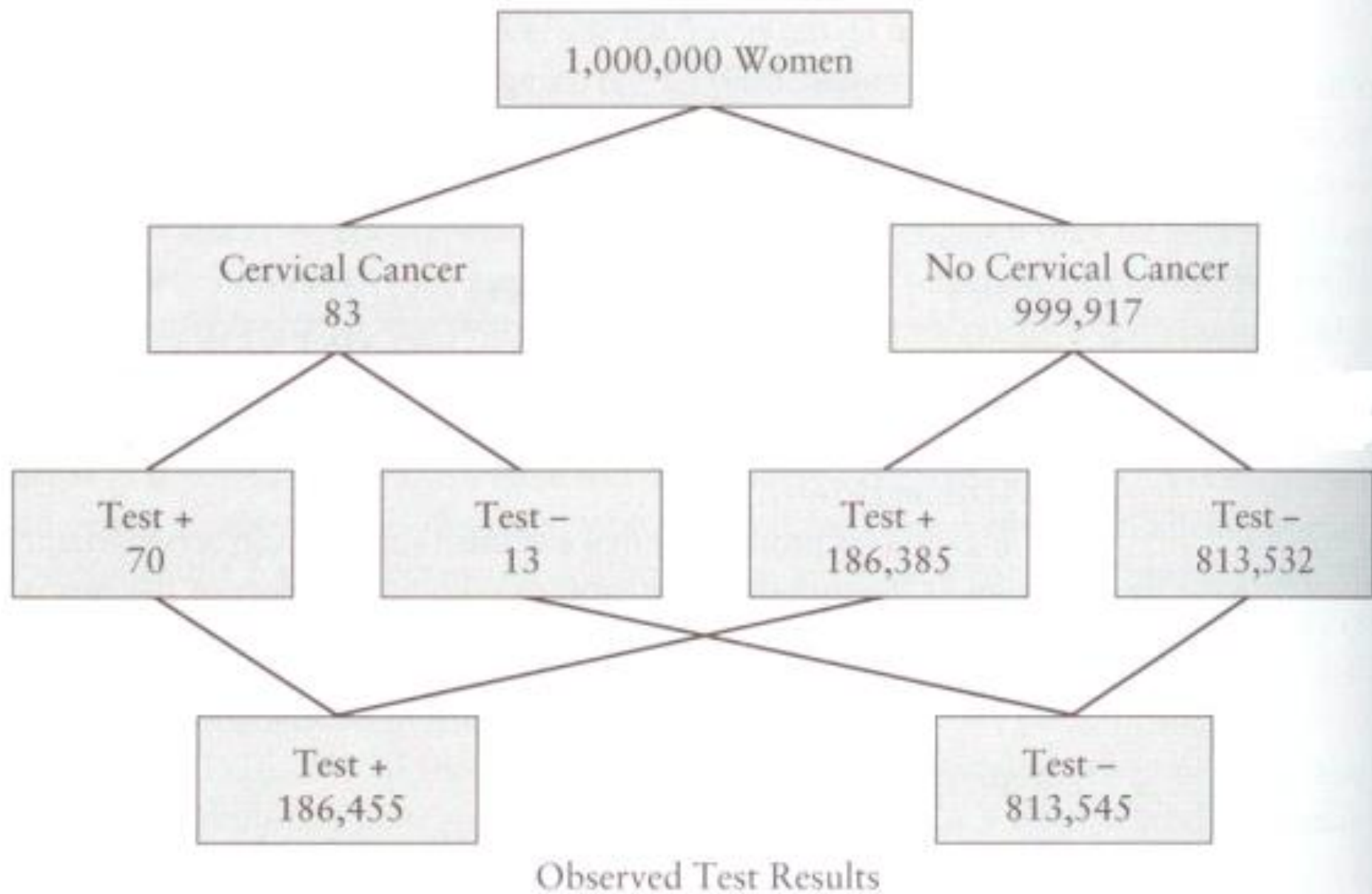


FIGURE 6.3

Performance of the Pap smear as a diagnostic test for cervical cancer

Cont'd

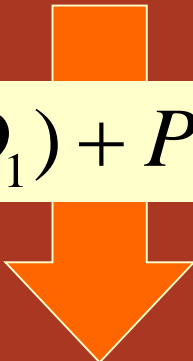
- Let D_1 represent the event that a woman has the disease cervical cancer, and D_2 the event that she does not.
- Let T^+ represent a positive Pap smear, and T^- be negative, accordingly.
- **We wish to compute $P(D_1|T^+)$** , that is, among those who tested positive, the probability of actually having cervical cancer.

Solution

$$P(A_1 \cup A_2 \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) = 1$$

$$P(A_i | B) = \frac{P(A_i)P(B | A_i)}{P(A_1)P(B | A_1) + \dots + P(A_n)P(B | A_n)}$$

General
formula
for Bayes'
theorem


$$P(D_1 \cup D_2) = P(D_1) + P(D_2) = 1$$

**D_1 and D_2 are
two exclusive
events.**

$$P(D_1 | T^+) = \frac{P(D_1)P(T^+ | D_1)}{P(D_1)P(T^+ | D_1) + P(D_2)P(T^+ | D_2)}$$

$$P(D_1 | T^+) = \frac{P(D_1)P(T^+ | D_1)}{P(D_1)P(T^+ | D_1) + P(D_2)P(T^+ | D_2)}$$

To get the answer, we need to first determine the three conditional probabilities in the red dotted boxes.

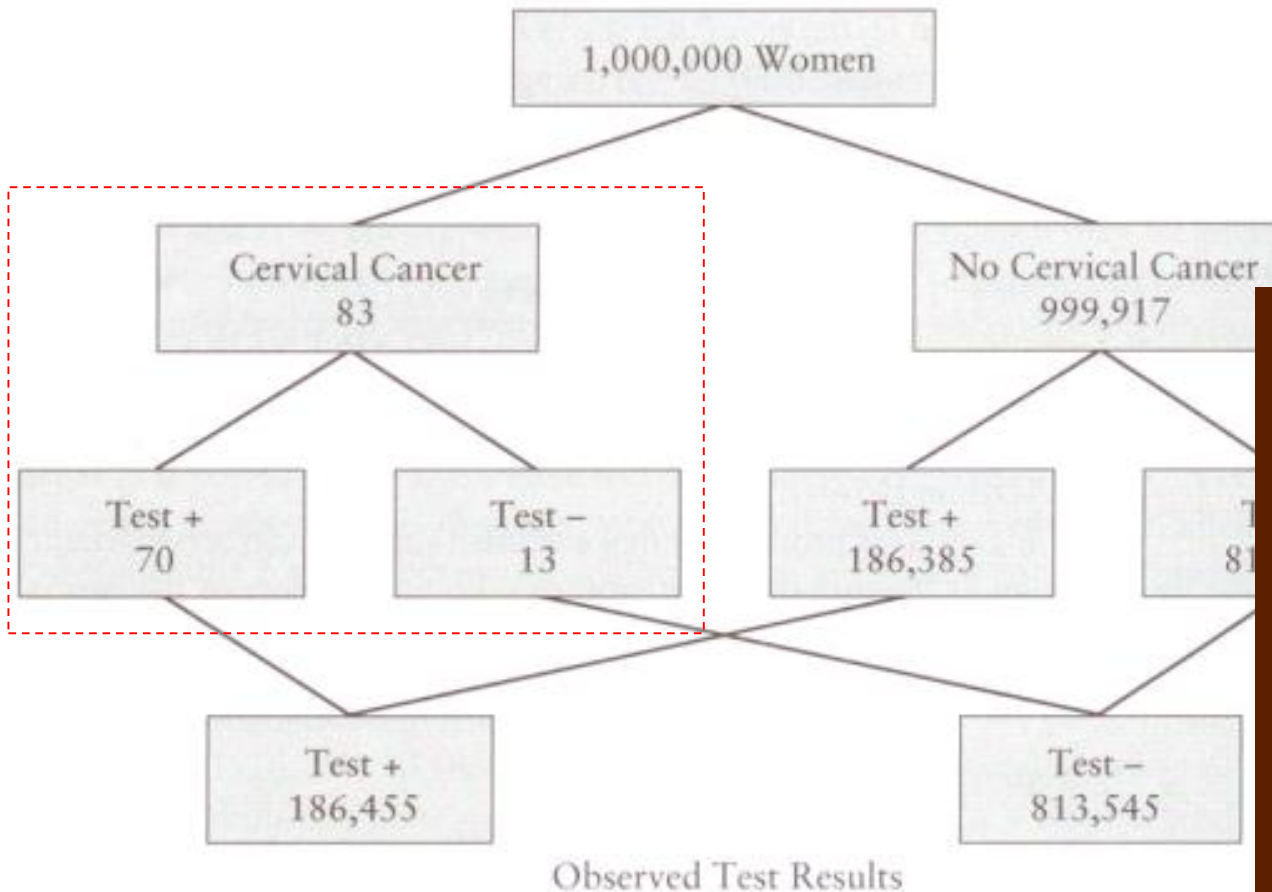


FIGURE 6.3

Performance of the Pap smear as a diagnostic test for cervical cancer

- It is easily see from the figure that $P(T^+|D_1) = 70 / 83 = 0.8434$.

$$P(D_1 | T^+) = \frac{P(D_1)P(T^+ | D_1)}{P(D_1)P(T^+ | D_1) + P(D_2)P(T^+ | D_2)}$$

- We also know that $P(T^+|D_2) = 186,385/999,917 = 0.1864$.

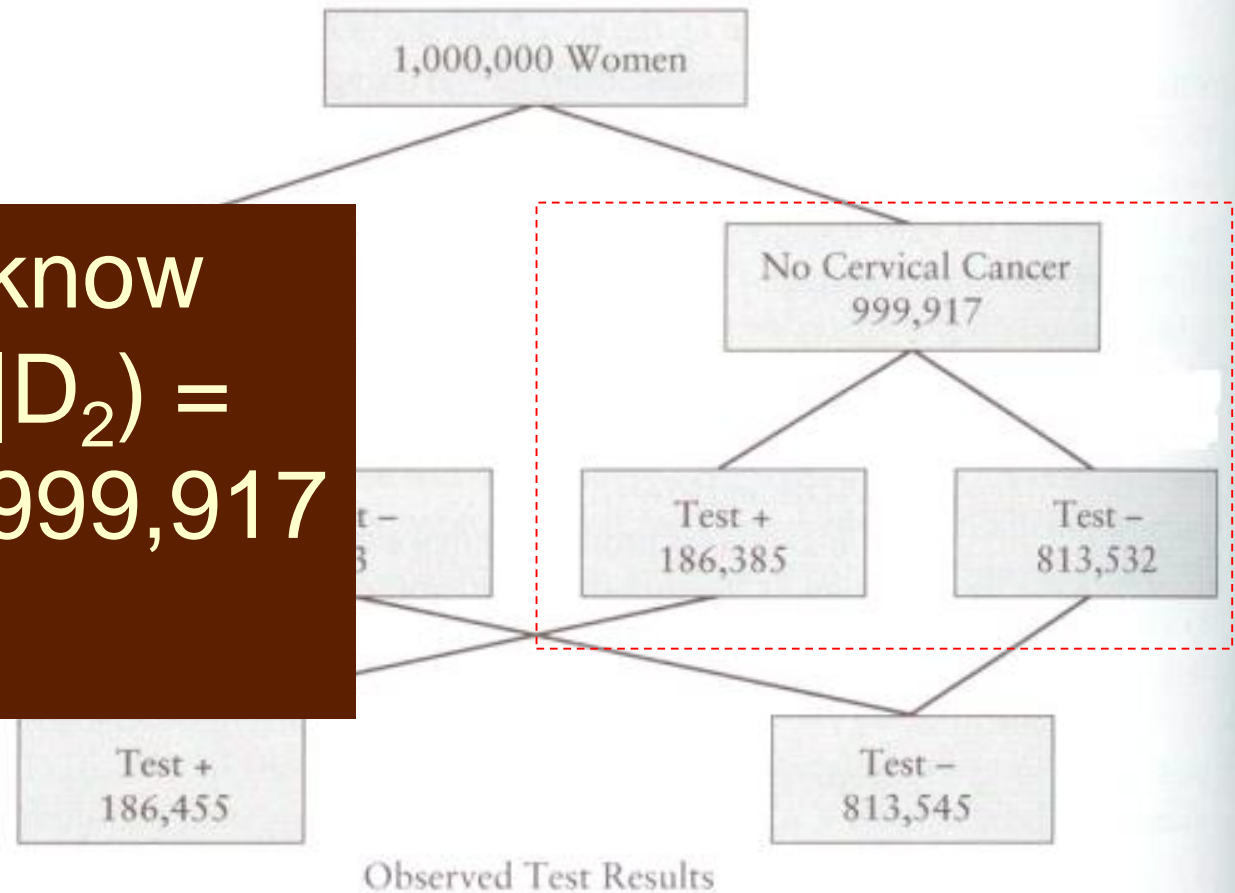


FIGURE 6.3
Performance of the Pap smear as a diagnostic test for cervical cancer

$$P(D_1 | T^+) = \frac{P(D_1)P(T^+ | D_1)}{P(D_1)P(T^+ | D_1) + P(D_2)P(T^+ | D_2)}$$

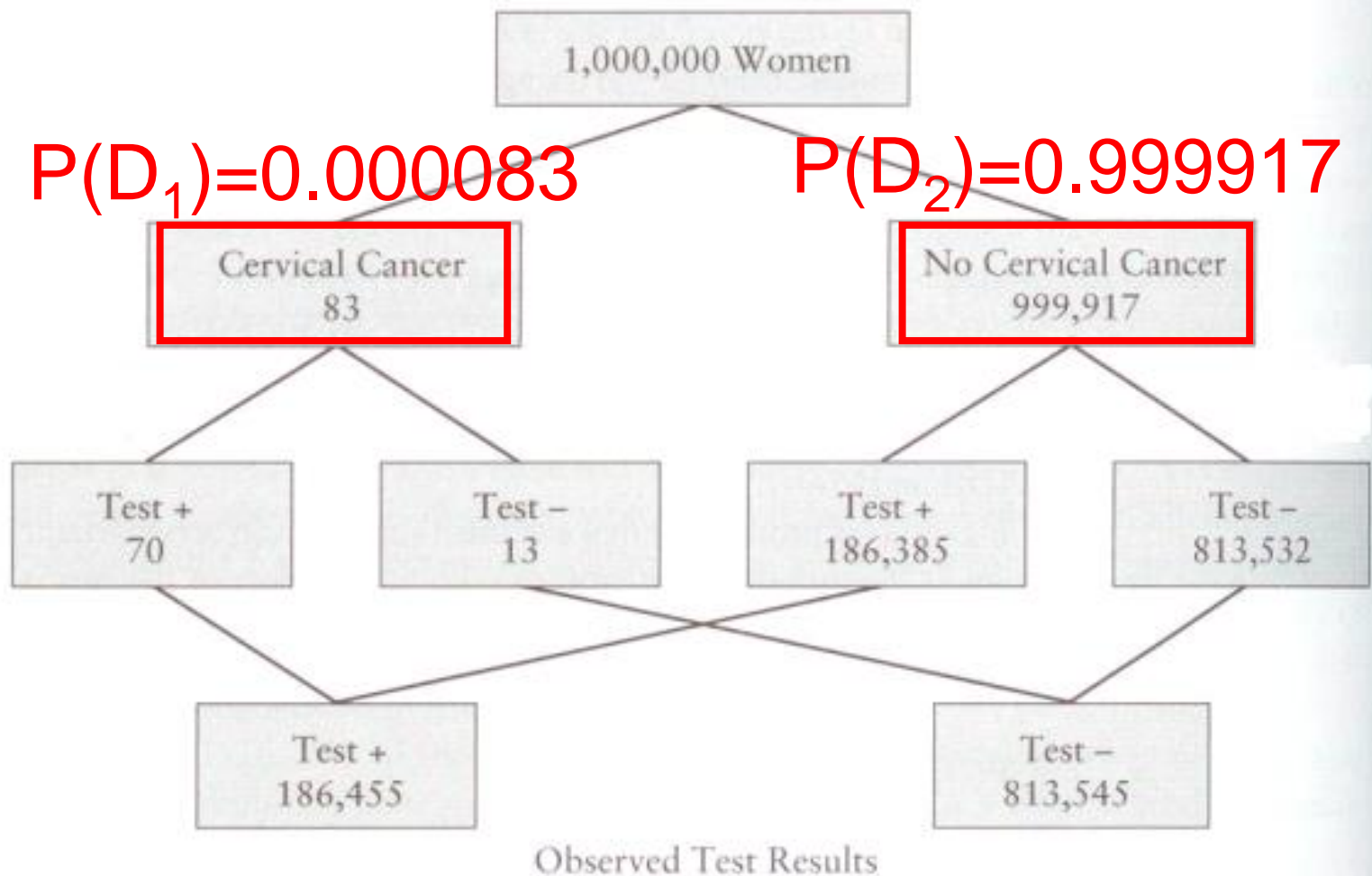


FIGURE 6.3

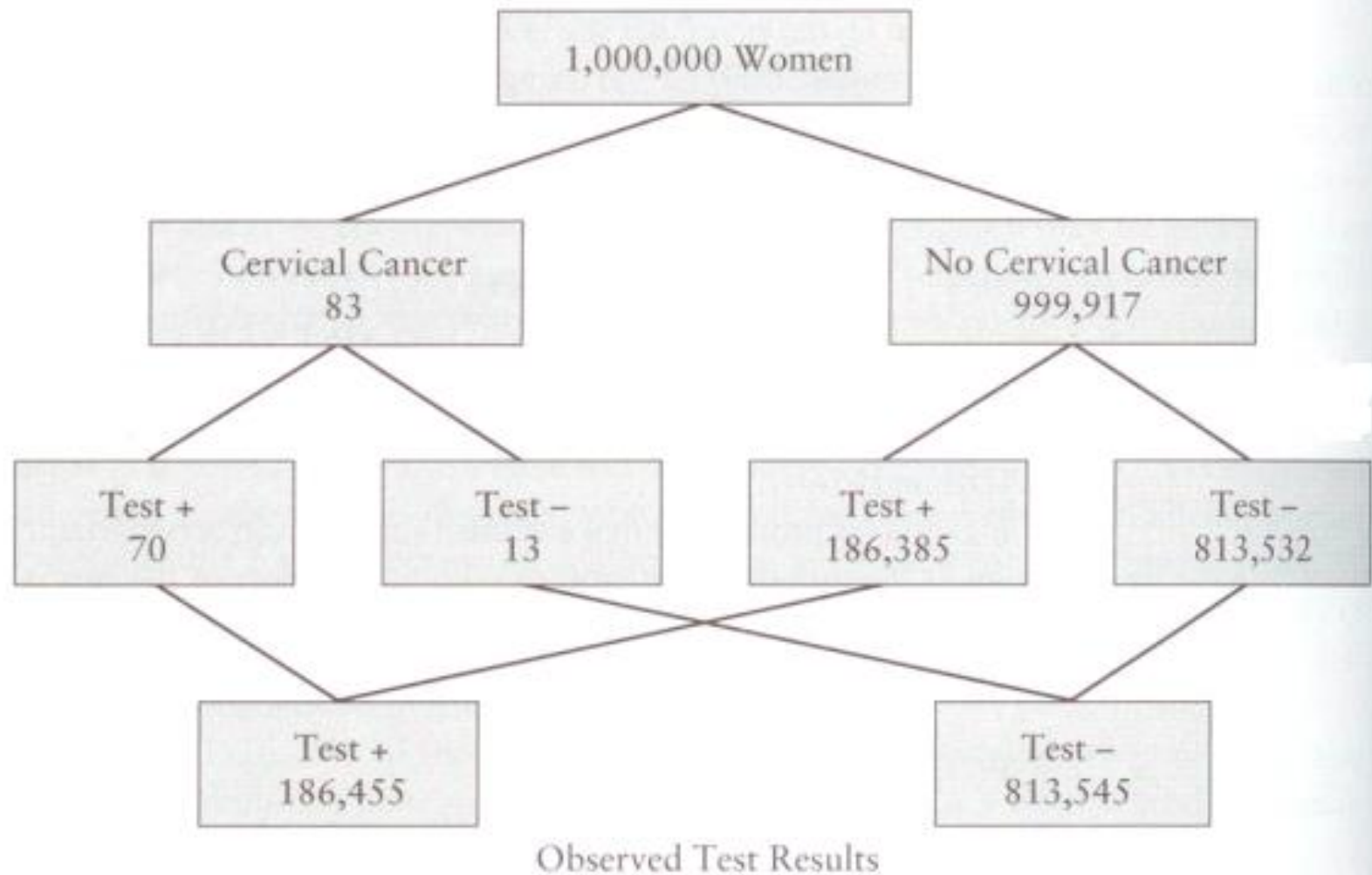
Performance of the Pap smear as a diagnostic test for cervical cancer

Finally, we have, according to the formula

$$\begin{aligned} P(D_1 | T^+) &= \frac{P(D_1)P(T^+ | D_1)}{P(D_1)P(T^+ | D_1) + P(D_2)P(T^+ | D_2)} \\ &= \frac{0.000083 \times 0.8434}{0.000083 \times 0.8434 + 0.999917 \times 0.1864} \\ &= 0.000375 \end{aligned}$$

That is, for every 1,000,000 positive Pap smears, only 375 represent true cases of cervical cancer. This is also called positive predictive value (PPV) that we have seen earlier.

☑ Can you compute the negative predictive value (NPV) for this case? That is, to compute $P(D_2 | T^-)$.



TP (patients who are correctly diagnosed as ill) = 70

TN (healthy ones got correctly diagnosed as healthy) = 813,532

FP (healthy ones got wrongfully diagnosed as ill) = 186,385

FN (patients got wrongfully diagnosed as healthy) = 13

TP (patients who are correctly diagnosed as ill) = 70

TN (healthy ones got correctly diagnosed as healthy) = 813,532

FP (healthy ones got wrongfully diagnosed as ill) = 186,385

FN (patients got wrongfully diagnosed as healthy) = 13

sensitivity

$$= \frac{TP}{TP + FN}$$

$$= \frac{70}{70 + 13}$$

$$= 84.34\%$$

specificity

$$= \frac{TN}{TN + FP}$$

$$= \frac{813532}{813532 + 186385}$$

$$= 81.36\%$$

Both values seem to be satisfactory!!!

Example # 4

- The level of serum ***creatinine*** (酩, found in blood) was used to test potential transplant rejection (器官移植排斥).
- When serum creatinine was greater than a prescribed threshold, the patient would be “diagnosed” to reject the organ.
- *High sensitivity* means to identify more patients who will actually reject the organ (less risk in actual transplant).

Sensitivity and specificity of serum creatinine level for predicting transplant rejection

| Serum Creatinine (mg %) | Sensitivity | Specificity |
|-------------------------|-------------|-------------|
| 1.2 | 0.939 | 0.123 |
| 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 | 0.364 | 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 |

- In this case, when the threshold was set to as low as 1.2, almost all transplant-rejecting patients would be identified (939 out of 1,000).
- When the threshold is set to a higher value (e.g., 2.9), we would miss ~70% of the rejecting cases (only 303 out of 1,000 are identified).

Sensitivity and specificity of serum creatinine level for predicting transplant rejection

| Serum Creatinine (mg %) | Sensitivity | Specificity |
|-------------------------|-------------|-------------|
| 1.2 | 0.939 | 0.123 |
| 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 | 0.364 | 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 |

- High sensitivity, however, corresponds to low specificity in this case.
- For example, a specificity of 0.123 means for 1,000 patients who won't actually reject the organ, we can only say 123 of them won't do so, and 877 of them would end up with false negative result, i.e., to reject the organ.

Sensitivity and specificity of serum creatinine level for predicting transplant rejection

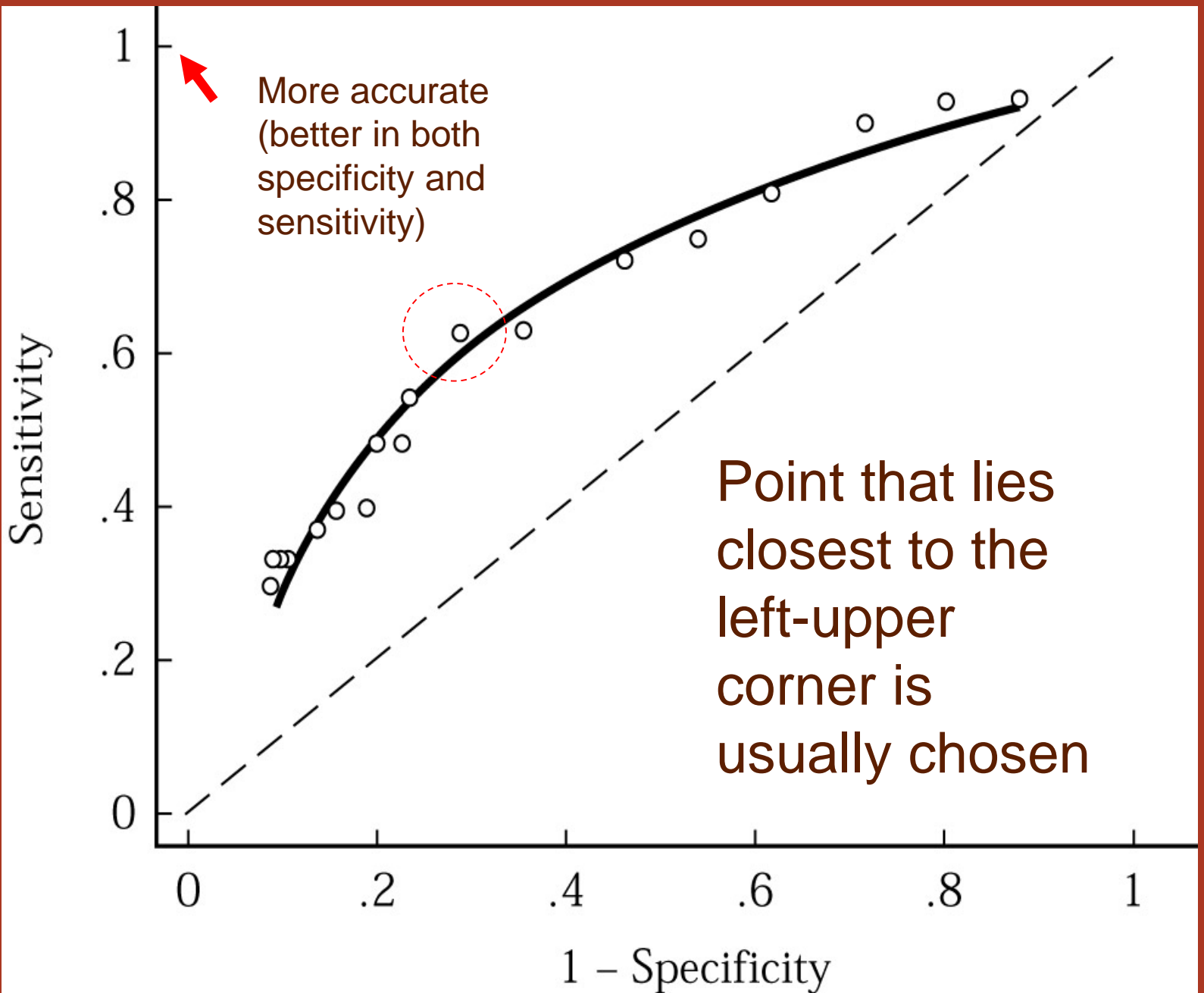
| Serum Creatinine (mg %) | Sensitivity | Specificity |
|-------------------------|-------------|-------------|
| 1.2 | 0.939 | 0.123 |
| 1.3 | 0.939 | 0.203 |
| 1.4 | 0.909 | 0.281 |
| 1.5 | 0.818 | 0.380 |
| 1.6 | 0.758 | 0.461 |
| 1.7 | 0.727 | 0.535 |
| 1.8 | 0.636 | 0.649 |
| 1.9 | 0.636 | 0.711 |
| 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 | 0.364 | 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 |



- There exists a trade-off between these two important parameters?
- How to locate one to maximize sensitivity as well as specificity?

6.4.3 ROC Curves

- The relationship between sensitivity and specificity may be illustrated by a graph known as a *Receive Operator Characteristic (ROC) curve (aka a Relative Operating Characteristic curve)*
- ROC curve is a line graph that plots the “sensitivity” against “1 – specificity” (or “one minus specificity”).



Sensitivity and specificity of serum creatinine level for predicting transplant rejection

| Serum Creatinine (mg %) | Sensitivity | Specificity |
|-------------------------|-------------|-------------|
| 1.2 | 0.939 | 0.123 |
| 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 | 0.364 | 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 |

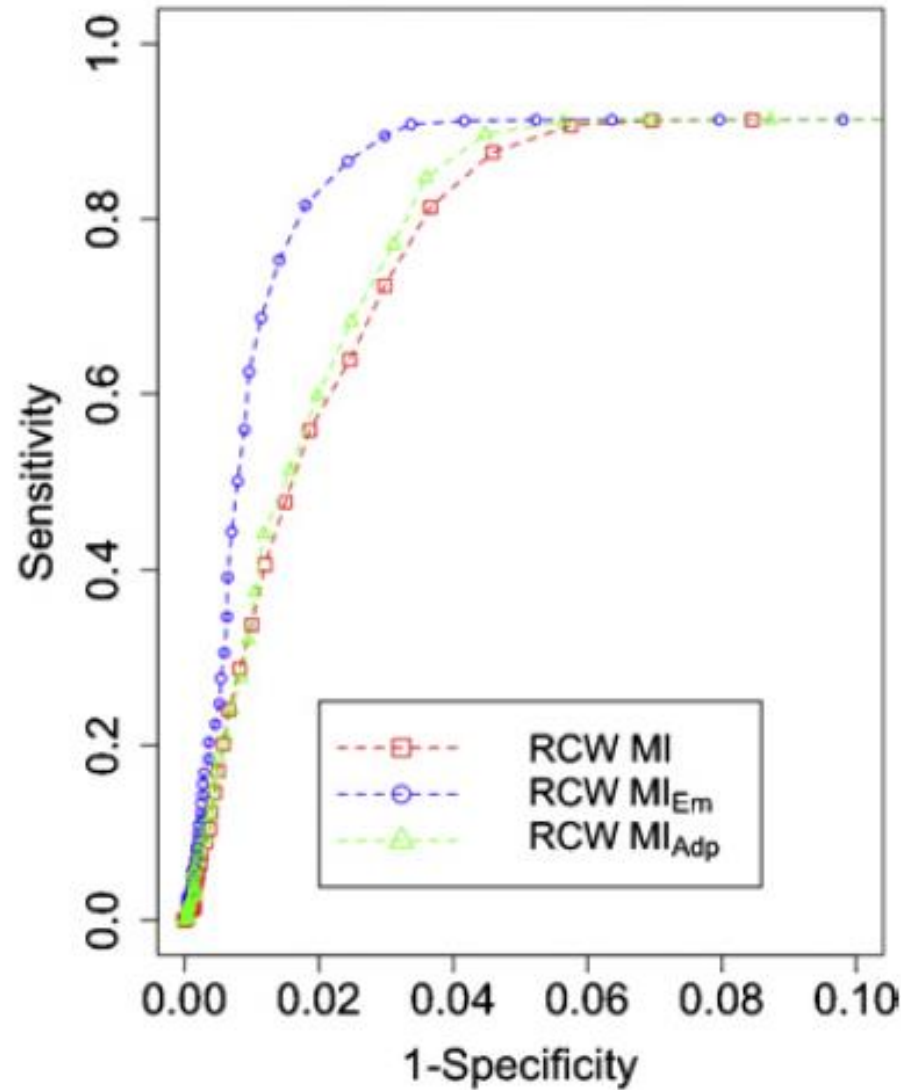
Left-upper-most point as revealed from an ROC curve (from previous slide).

More on ROC curves (wiki)

- ROC analysis provides tools to select possibly optimal models and to discard suboptimal ones independently from (and prior to specifying) the cost context or the class distribution.
- ROC analysis is related in a direct and natural way to cost/benefit analysis of diagnostic decision making.

Cont'd

- The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields and was soon introduced to psychology to account for perceptual (知覺的) detection of stimuli.
- ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research.



For different curves, the one giving the maximum area under curve (AUC) is usually chosen as the optimal method under investigation.

6.4.4 Prevalence evaluation

- Prevalence (盛行率) or prevalence proportion, in epidemiology, is the proportion of a population found to have a condition (typically a disease or a risk factor such as smoking or seat-belt use).
- It is usually expressed as a fraction, as a percentage or as the number of cases per 10,000 or 100,000 people.

Example #5

- A program was conducted to screen ***HIV infections in mothers.*** (The purpose is to know whether a mother is infected or not.)
- One cannot test on mothers. They may not like to be tested at all.
- Instead, one can test naturally on newborn babies to understand infections in mothers.

Example #5 – cont'd

- Since maternal antibodies (抗體) cross the placenta (胎盤), the presence of antibodies in an infant signals infection in the mother.
- No verification of the results is possible. (No mother is tested!!!)
- ***Is the result from testing newborns really represent HIV prevalence in mothers?***

Defining various events

- H : the event that a mother is infected with HIV.
- H^C : the event that a mother is NOT infected with HIV.
- n : total number of infants tested
- n^+ : number of infants with positive results
- T^+ : the event for a positive test result for an infant
- T^- : the event for a negative test result for an infant



CANADA

ONTARIO

Lake Ontario

Niagara

Orleans

Monroe

Wayne

Genesee

Erie

Wyoming

Livingston

Ontario

Yates

Saratoga

Cayuga

Onondaga

Madison

Cortland

Chenango

Otsego

Montgomery

Schenectady

Schoharie

Albany

Greene

Columbia

Ulster

Dutchess

Orange

Putnam

Rockland

Westchester

Bronx

New York

Nassau

Kings

Queens

Richmond

MASSACHUSETTS

CONNECTICUT

PENNSYLVANIA

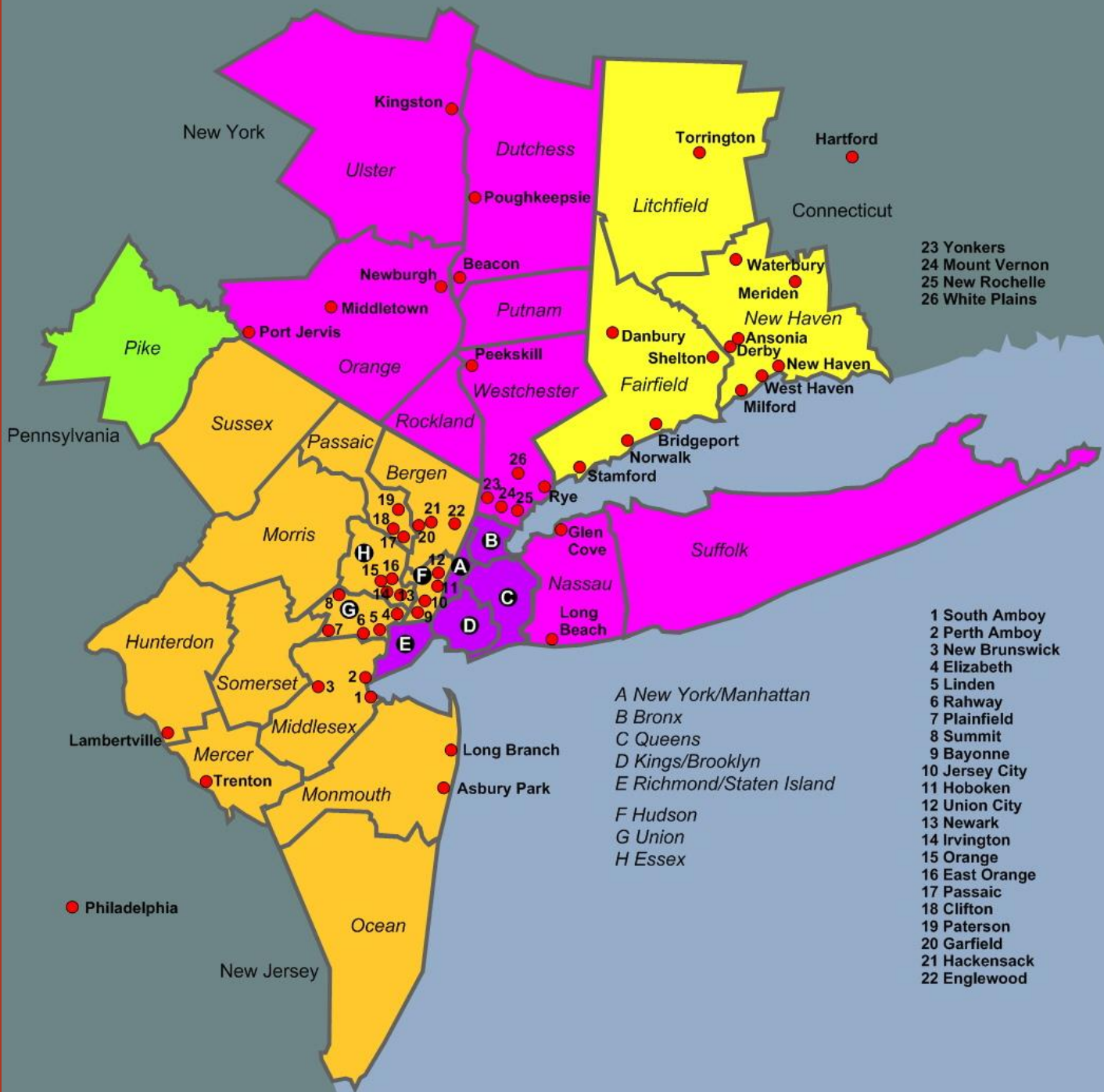
NEW JERSEY

QUÉBEC

VERMONT

NEW HAMPSHIRE

ATLANTIC OCEAN



- 23 Yonkers
- 24 Mount Vernon
- 25 New Rochelle
- 26 White Plains

- A New York/Manhattan
- B Bronx
- C Queens
- D Kings/Brooklyn
- E Richmond/Staten Island
- F Hudson
- G Union
- H Essex

- 1 South Amboy
- 2 Perth Amboy
- 3 New Brunswick
- 4 Elizabeth
- 5 Linden
- 6 Rahway
- 7 Plainfield
- 8 Summit
- 9 Bayonne
- 10 Jersey City
- 11 Hoboken
- 12 Union City
- 13 Newark
- 14 Irvington
- 15 Orange
- 16 East Orange
- 17 Passaic
- 18 Clifton
- 19 Paterson
- 20 Garfield
- 21 Hackensack
- 22 Englewood

New York City



TABLE 6.2

Percentage of **HIV-positive newborns** by region for the state of New York,
December 1987–March 1990

| Region | Number Positive | Total Tested | Percent Positive |
|---------------------------------|--------------------|-----------------|---------------------|
| New York State exclusive of NYC | 601 | 346,522 | 0.17 |
| NYC suburban | 329 | 120,422 | 0.27 |
| Mid-Hudson Valley | 71 | 29,450 | 0.24 |
| Upstate urban | 119 | 88,088 | 0.14 |
| Upstate rural | 82 | 108,562 | 0.08 |
| New York City | 3650 | 294,062 | 1.24 |
| Manhattan | 799 | 50,364 | 1.59 |
| Bronx | 998 | 58,003 | 1.72 |
| Brooklyn | 1352 | 104,613 | 1.29 |
| Queens | 424 | 67,474 | 0.63 |
| Staten Island | 77 | 13,608 | 0.57 |

Taking Manhattan for example

- $n = 50,364$ infants were tested and $n^+ = 799$ were positive, that is:

$$\frac{n^+}{n} = \frac{799}{50,364} = 0.0159.$$

- In other words, $P(T^+) = 0.0159$ (from infants)
- We want to know $P(H)$: the prevalence of mother infection.
- **Is it true $P(H) = P(T^+) = 0.0159$?**

A test is not perfect...

- If the screen tests were perfect, then $P(H) = P(T^+) = 0.0159$.
- **We may have both true positive and false positive cases.**
- **Similarly, we may have both true negative and false negative.**

- Recall earlier we defined conditional probability as:

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

given that $P(A) \neq 0$

- Similarly, we may have

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

given that $P(B) \neq 0$

(and is read "the conditional probability of B , given A " or "the probability of B under the condition A ".)

Infants tested positive came from two sources:

Mother infected (H), and infant tested positive (T^+) : true positive

Mother not infected (H^C), and infant tested positive (T^+) : false positive

$$\begin{aligned}P(T^+) &= P(T^+ \cap H) + P(T^+ \cap H^C) \\&= P(T^+ | H)P(H) + P(T^+ | H^C)P(H^C) \\&= P(T^+ | H)P(H) + P(T^+ | H^C)[1 - P(H)]\end{aligned}$$

(from previous page)

$$\begin{aligned}P(T^+) &= P(T^+ \cap H) + P(T^+ \cap H^C) \\&= P(T^+ | H)P(H) + P(T^+ | H^C)P(H^C) \\&= P(T^+ | H)P(H) + \boxed{P(T^+ | H^C)[1 - P(H)]}\end{aligned}$$

$$P(T^+) = P(T^+ | H) \textcircled{P(H)} + \boxed{P(T^+ | H^C) - P(T^+ | H^C) \textcircled{P(H)}}$$

$$P(T^+) = P(H)[P(T^+ | H) - P(T^+ | H^C)] + P(T^+ | H^C)$$

$$P(T^+) - P(T^+ | H^C) = P(H)[P(T^+ | H) - P(T^+ | H^C)]$$

- Solving the previous equation for $P(H)$ leads to:

$$\begin{aligned} P(H) &= \frac{P(T^+) - P(T^+ | H^c)}{P(T^+ | H) - P(T^+ | H^c)} \\ &= \frac{(n^+/n) - P(T^+ | H^c)}{P(T^+ | H) - P(T^+ | H^c)}. \end{aligned}$$

- $P(T^+ | H)$: Those infected mothers being tested positive in infants. **This is the sensitivity of the test.**
- $P(T^+ | H^c) = 1 - P(T^- | H^c)$, the last term represents healthy mothers being tested negative in infants. **This is the specificity of the test.**

- Assuming that this test has 0.99 sensitivity and 0.998 specificity:

$$\begin{aligned} P(H) &= \frac{P(T^+) - P(T^+ | H^c)}{P(T^+ | H) - P(T^+ | H^c)} \\ &= \frac{(n^+/n) - P(T^+ | H^c)}{P(T^+ | H) - P(T^+ | H^c)}. \end{aligned}$$

$$\begin{aligned} P(H) &= \frac{0.0159 - (1 - 0.998)}{0.99 - (1 - 0.998)} \\ &= 0.0141, \end{aligned}$$

- A scale-down from 0.0159 to 0.0141.

Consider a different region...

TABLE 6.2

Percentage of HIV-positive newborns by region for the state of New York, December 1987–March 1990

| Region | Number Positive | Total Tested | Percent Positive |
|---------------------------------|-----------------|--------------|------------------|
| New York State exclusive of NYC | 601 | 346,522 | 0.17 |
| NYC suburban | 329 | 120,422 | 0.27 |
| Mid-Hudson Valley | 71 | 29,450 | 0.24 |
| Upstate urban | 119 | 88,088 | 0.14 |
| Upstate rural | 82 | 108,562 | 0.08 |
| New York City | 3650 | 294,062 | 1.24 |
| Manhattan | 799 | 50,364 | 1.59 |
| Bronx | 998 | 58,003 | 1.72 |
| Brooklyn | 1352 | 104,613 | 1.29 |
| Queens | 424 | 67,474 | 0.63 |
| Staten Island | 77 | 13,608 | 0.57 |

- In upstate urban region of New York:

$$\frac{n^+}{n} = \frac{119}{88,088} \\ = 0.0014,$$

- By using the same formula, we have:

$$P(H) = \frac{0.0014 - (1 - 0.998)}{0.99 - (1 - 0.998)} \\ = -0.0006.$$

- **This however, turns into a negative prevalence? [Making no sense!!!]**

A brief summary

- Note that $P(T^+) = 0.0014$ in the second case, which is very small comparing with 0.0159 in the first case.
- **The testing procedure is not accurate enough to measure the very low prevalence of HIV in the second case.**