

Medical Statistics and Data Analysis

Lecture 5: Probability

Zhiyu Wan, PhD (wanzhy@shanghaitech.edu.cn)

Assistant Professor of Biomedical Engineering

ShanghaiTech University

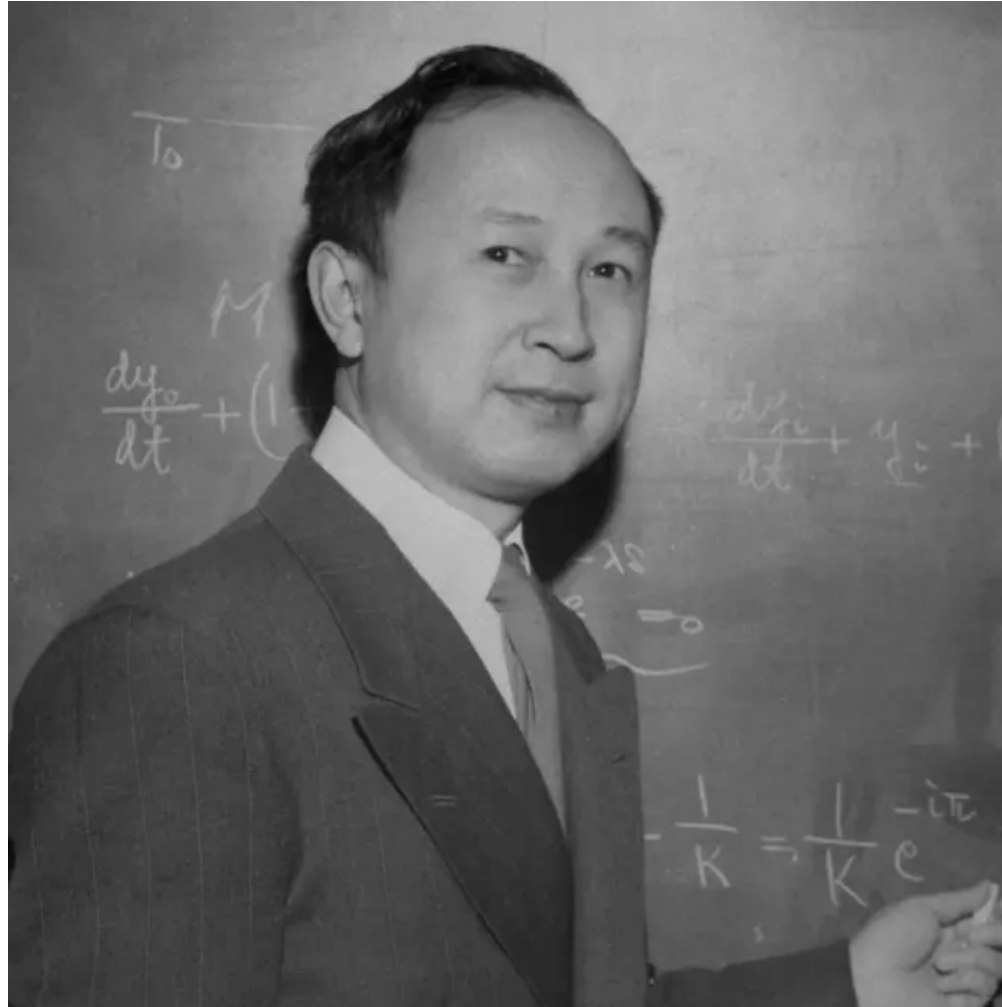
October 10, 2026

Alexandr Wang

- Co-founded Scale AI, 2016
- Chief AI Officer, Meta, 2025



Qian Xuesen (钱学森)



Learning Goals of This Lecture

- Bayesian inference.
- Receiver operating characteristic (ROC) curves.

Introduction

- In addition to describing data, we might want to test specific inferences about the behavior of data.
- Hypothesis: Women who have their first child after the age of 30 are more likely to develop breast cancer than those who have their first child before age 20.
- Sample size: 2000 women; 45-54 years of age; of which, 1000 had their first child before age 20 and 1000 after age 30. This is a limited sample size and results may not be conclusive.
- The sample size may be increased to 10,000 women; however, the apparent difference in rate of breast cancer occurrence may still be due to chance.
- To set a framework for evaluating occurrence, we introduce the concept of **probability**.

Probability of a Male Livebirth During the Period 1965–1974

Time period	Number of male livebirths (a)	Total number of livebirths (b)	Empirical probability of a male livebirth (a/b)
1965	1,927,054	3,760,358	0.51247
1965–1969	9,219,202	17,989,361	0.51248
1965–1974	17,857,857	34,832,051	0.51268

- The probability of live male births in 1965 was 0.51247; in 1965–69 was 0.51248; in 1965–74 was 0.51268.
- These **empirical probabilities** are based on finite amount of data.
- The sample size could be expanded indefinitely and an increasingly more precise estimate of the probability obtained.

- The **sample space** is the set of all possible outcomes.
- An **event** is any set of outcomes of interest.
- The **probability** of an event is the relative frequency of this set of outcomes over an indefinitely large (or infinite) number of trials.
- In real life, experiments cannot be performed an infinite number of times. Hence, probabilities are estimated from empirical probabilities obtained from larger samples.
- Theoretical probability models may also be constructed from which probabilities of many different kinds of events can be computed.
- Comparing empirical probabilities with theoretical probabilities enables us to assess the **goodness-of-fit** of probability models.

- The probability of an event E , denoted by $\Pr(E)$, always satisfies
$$0 \leq \Pr(E) \leq 1$$
- If outcomes A and B are two events that cannot both happen at the same time, then

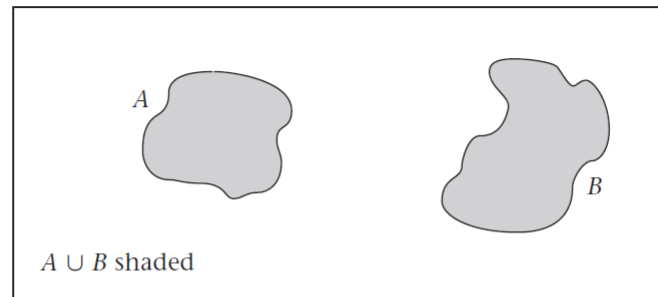
$$\Pr(A \text{ or } B \text{ occurs}) = \Pr(A) + \Pr(B)$$

- Example:
- Let A = person has normotensive diastolic blood pressure (DBP) ($\text{DBP} < 80$).
- B = person has borderline DBP readings ($80 \leq \text{DBP} < 95$).
- $\Pr(A) = 0.7$ and $\Pr(B) = 0.1$.
- Z = event that person has $\text{DBP} < 95$.
- Then, $\Pr(Z) = \Pr(A) + \Pr(B) = 0.8$
- The two events A and B are **mutually exclusive** if they cannot both happen at the same time.

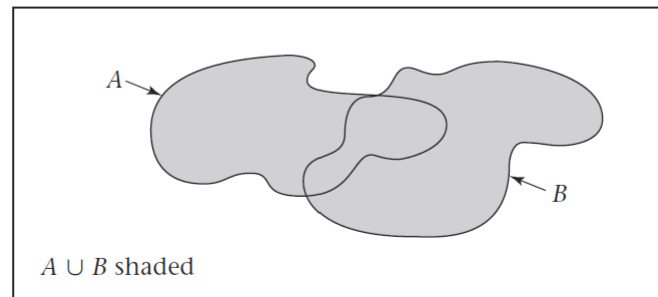
Some Useful Probabilistic Notations

- The symbol $\{ \}$ is used as shorthand for the phrase “the event.”
- $A \cup B$ is the event that either A or B occurs, or they both occur.

FIGURE 3.1 Diagrammatic representation of $A \cup B$: (a) A, B mutually exclusive; (b) A, B not mutually exclusive



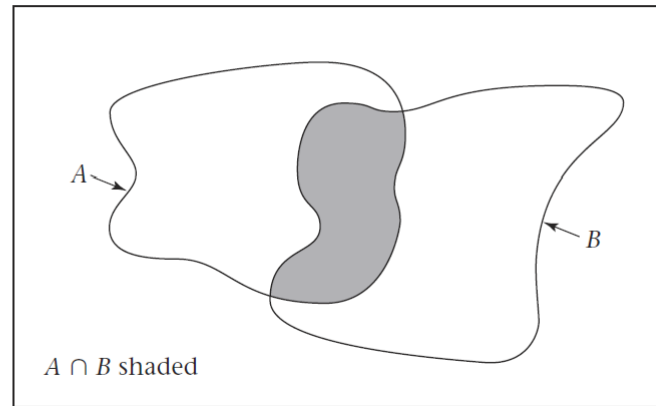
(a)



(b)

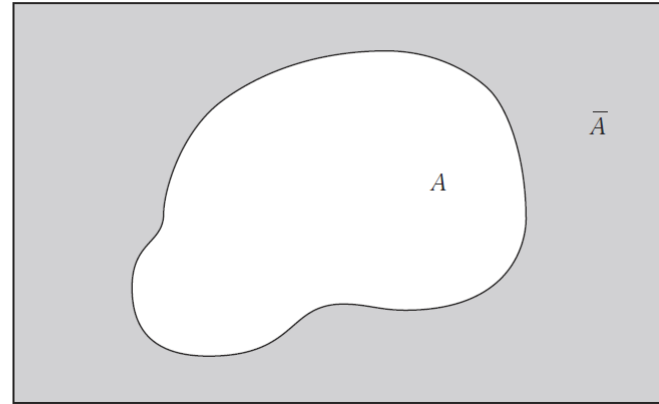
- $A \cap B$ is the event that both A and B occur simultaneously.

FIGURE 3.2 Diagrammatic representation of $A \cap B$



- $\bar{A}$ is the event that A does not occur. It is called the **complement** of A . Notice that $\Pr(\bar{A}) = 1 - \Pr(A)$, because $\bar{A}$ occurs only when A does not occur.

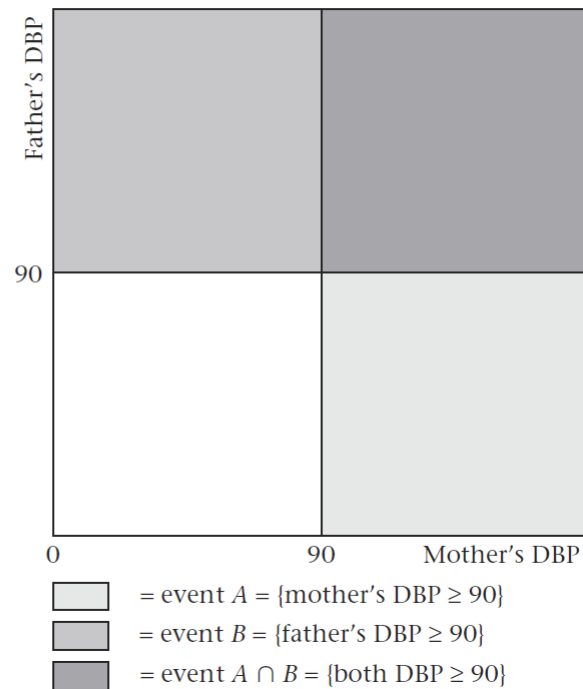
FIGURE 3.3 Diagrammatic representation of $\bar{A}$



- Example:
- If A and C are two events:
- $A = \{X < 90\}$; $C = \{X \geq 90\}$
- Then, $C = \bar{A}$, because C can only occur when A does not occur.

Multiplication Law of Probability

FIGURE 3.4 Possible diastolic blood-pressure measurements of the mother and father within a given family



- Events A and B are called **independent** events if
- $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$
- If $\Pr(A) = 0.1$ and $\Pr(B) = 0.2$,
- Probability that both mother and father are hypertensive,
- $\Pr(A \cap B) = \Pr(A) \times \Pr(B) = 0.1 \times 0.2 = 0.02$
- If $A_1, \dots, A_k$ are mutually independent events, then $\Pr(A_1 \cap A_2 \cap \dots \cap A_k) = \Pr(A_1) \times \Pr(A_2) \times \dots \times \Pr(A_k)$

- Two events A, B are **dependent** if
$$Pr(A \cap B) \neq Pr(A) \times Pr(B)$$
- For a continuous random variable, the probability of exactly obtaining any exact value is 0.
- However, there exists the notion that certain ranges of values occur more frequently than others.
- This notion can be quantified using the concept of a **probability-density function (pdf)**.

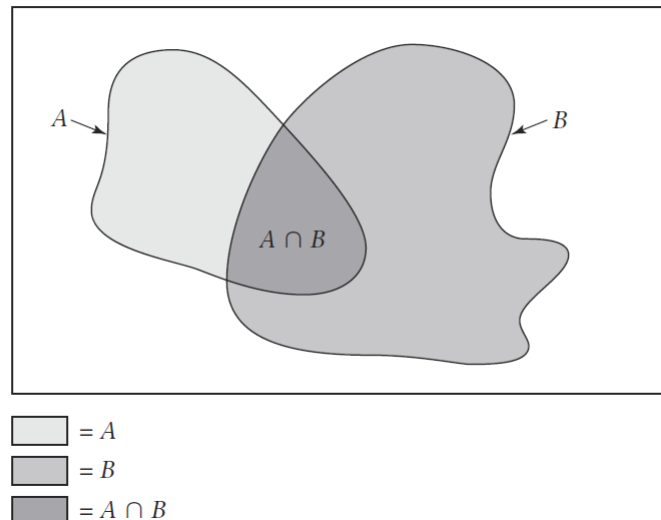
Addition Law of Probability

- If A and B are any events, then

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

- Special case: If events A and B are mutually exclusive, then $Pr(A \cap B) = 0$;
And the addition law reduces to $Pr(A \cup B) = Pr(A) + Pr(B)$

FIGURE 3.5 Diagrammatic representation of the addition law of probability



- For independent events A and B :

$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)]$$

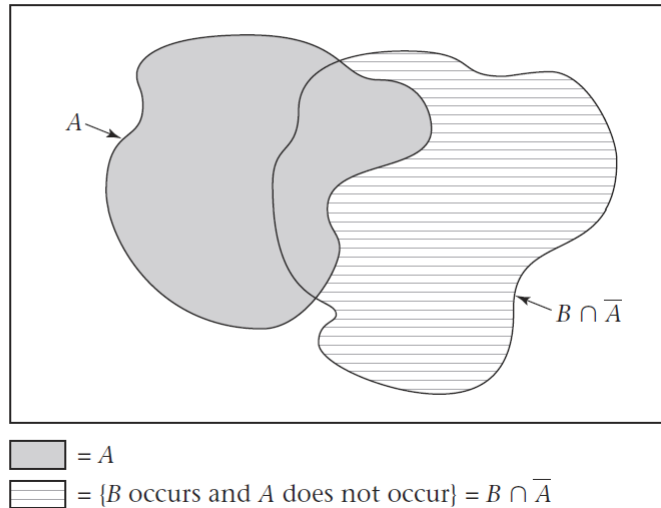
- Two mutually exclusive events: {A occurs} and {B occurs and A does not occur}
- Similarly, for three events A , B , and C :

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C)$$

$$-Pr(A \cap B) - Pr(A \cap C)$$

$$-Pr(B \cap C) + Pr(A \cap B \cap C)$$

FIGURE 3.6 Diagrammatic representation of the addition law of probability for independent events



Conditional Probability

- The quantity $Pr(A \cap B)/Pr(A)$ is defined as the **conditional probability of B given A** , which is written $Pr(B | A)$.
- If A and B are independent events, then $Pr(B | A) = Pr(B) = Pr(B | \bar{A})$
- If two events A, B are dependent, then $Pr(B | A) \neq Pr(B) \neq Pr(B | \bar{A})$ and $Pr(A \cap B) \neq Pr(A) \times Pr(B)$
- The **relative risk (RR)** of B given A is

$$RR = \frac{Pr(B | A)}{Pr(B | \bar{A})}$$

- If A and B are independent, then RR is 1.
- If A and B are dependent, then RR is different from 1.
- The more the dependence between events increases, the further the RR will be from 1.

Total-Probability Rule

- For any events A and B , $Pr(B) = Pr(B | A) \times Pr(A) + Pr(B | \bar{A}) \times Pr(\bar{A})$
- If event B occurs, it must occur either with A or without A .

$$Pr(B) = Pr(B \cap A) + Pr(B \cap \bar{A})$$

- From definition of conditional probability,

$$Pr(B \cap A) = Pr(A) \times Pr(B | A)$$

- and

$$Pr(B \cap \bar{A}) = Pr(\bar{A}) \times Pr(B | \bar{A})$$

- On substitution,

$$Pr(B) = Pr(B | A)Pr(A) + Pr(B | \bar{A})Pr(\bar{A})$$

- The unconditional probability of B is the sum of the conditional probability of B given A times the unconditional probability of A plus the conditional probability of B given A not occurring times the unconditional probability of A not occurring.

- The probability of B is expressed in terms of 2 mutually exclusive events A and $\bar{A}$. In many instances, the probability of B will need to be expressed in terms of more than two mutually exclusive events, $A_1, A_2, \dots, A_k$.
- A set of events $A_1, \dots, A_k$ is **exhaustive** if at least one of the events must occur.
- Let $A_1, \dots, A_k$ be **mutually exclusive and exhaustive events**. The unconditional probability of B ($Pr(B)$) can then be written as a weighted average of the conditional probabilities of B given A_i ($Pr(B | A_i)$) as follows:

$$Pr(B) = \sum_{i=1}^k Pr(B | A_i) \times Pr(A_i)$$

- **Generalized Multiplication Law of Probability**
- If $A_1, \dots, A_k$ are an arbitrary set of events, then

$$\begin{aligned} Pr(A_1 \cap A_2 \cap \dots \cap A_k) &= Pr(A_1) \\ &\quad \times Pr(A_2 | A_1) \\ &\quad \times Pr(A_3 | A_2 \cap A_1) \\ &\quad \times \dots \\ &\quad \times Pr(A_k | A_{k-1} \cap \dots \cap A_2 \cap A_1) \end{aligned}$$

Bayes' Rule and Screening Tests

- The **predictive value positive (PV⁺)** or **positive predictive value (PPV)** of a screening test is the probability that a person has a disease given that the test is positive.

$$Pr(\text{disease} \mid \text{test}^+) = Pr(D^+ \mid T^+)$$

- The **predictive value negative (PV⁻)** or **negative predictive value (NPV)** of a screening test is the probability that a person does not have a disease given that the test is negative.

$$Pr(\text{no disease} \mid \text{test}^-) = Pr(D^- \mid T^-)$$

- A symptom or set of symptoms can be regarded as a screening test for a disease. The higher the PV of the symptoms, the more valuable the test will be.
- Clinicians often cannot directly measure the PV of a set of symptoms. However, they can measure how often specific symptoms occur in diseased and normal people.
- **True positive (真阳性, TP):** Positive test result when the disease or condition being tested for is actually present.
- **True negative (真阴性, TN):** Negative test result when the disease or condition being tested for is actually absent.
- **False negative (假阴性, FN):** Negative test result when the disease or condition being tested for is actually present.
- **False positive (假阳性, FP):** Positive test result when the disease or condition being tested for is actually absent.

- **Sensitivity (灵敏度) or True positive rate (真阳性率, TPR)** of a symptom (or a set of symptoms): the probability that the symptom is present given that the person has a disease.

$$\text{Sensitivity} = Pr(T^+ | D^+)$$

- **Specificity (特异度) or True negative rate (真阴性率, TNR)** of a symptom (or a set of symptoms): the probability that the symptom is not present given that the person does not have a disease.

$$\text{Specificity} = Pr(T^- | D^-)$$

- **False negative rate (假阴性率, FNR)** of a symptom (or a set of symptoms): The probability that the symptom is not present given that the person has a disease.

- $FNR = Pr(T^- | D^+) = 1 - \text{Sensitivity}$

- **False positive rate (假阳性率, FPR)** of a symptom (or a set of symptoms): The probability that the symptom is present given that the person does not have a disease.

- $FPR = Pr(T^+ | D^-) = 1 - \text{Specificity}$

- For a symptom to be effective in predicting a disease, the **specificity** and **sensitivity** must be high.
- The level of **prostate-specific antigen (PSA)** in the blood is frequently used as a screening test for prostate cancer.

TABLE 3.2 Association between PSA and prostate cancer

PSA test result	Prostate cancer	Frequency
+	+	92
+	-	27
-	+	46
-	-	72

Contingency Table or Confusion Matrix

	T+	T-	
D+	TP = 92	FN = 46	138
D-	FP = 27	TN = 72	99
	119	118	237

$$\text{Sensitivity} = Pr(T^+ | D^+) = TP / (TP + FN) = 92 / (92 + 46) = 66.67\%$$

$$\text{Specificity} = Pr(T^- | D^-) = TN / (TN + FP) = 72 / (72 + 27) = 72.73\%$$

$$FNR = Pr(T^- | D^+) = FN / (TP + FN) = 46 / (92 + 46) = 33.33\%$$

$$FPR = Pr(T^+ | D^-) = FP / (TN + FP) = 27 / (72 + 27) = 27.27\%$$

$$PV^+ = Pr(D^+ | T^+) = TP / (TP + FP) = 92 / (92 + 27) = 77.31\%$$

$$PV^- = Pr(D^- | T^-) = TN / (TN + FN) = 72 / (72 + 46) = 61.02\%$$

Bayes' Rule

- Let A = symptom and B = disease.

$$PV^+ = Pr(B | A) = \frac{Pr(A | B) \times Pr(B)}{Pr(A | B) \times Pr(B) + Pr(A | \bar{B}) \times Pr(\bar{B})}$$

- In words, this can be written as:

$$PV^+ = \frac{\text{Sensitivity} \times x}{\text{Sensitivity} \times x + (1 - \text{Specificity}) \times (1 - x)}$$

- where $x = Pr(B)$ = prevalence of disease in the reference population.
- Similarly,

$$PV^- = \frac{\text{Specificity} \times (1 - x)}{\text{Specificity} \times (1 - x) + (1 - \text{Sensitivity}) \times x}$$

Generalized Bayes' Rule

- Let $B_1, B_2, \dots, B_k$ be a set of mutually exclusive and exhaustive disease states. Let A be a symptom or a set of symptoms.

$$Pr(B_i | A) = \frac{Pr(A | B_i) \times Pr(B_i)}{\sum_{j=1}^k Pr(A | B_j) \times Pr(B_j)}$$

Bayesian Inference

- It is an alternative definition of probability and inference.
- It “rejects” the idea of the definition of probability sometimes called the **frequency definition of probability** (a theoretical concept).
- Bayesians conceive of two types of probability:
 - **Prior probability** of an event is the best guess by the observer of an event's likelihood in the absence of *data*. This may be a single number, or a range of likely values, perhaps with weights attached to each possible value.
 - **Posterior probability** of an event is the likelihood that an event will occur after collecting some empirical *data*. It is obtained by integrating information from the prior probability with additional data related to the event in question.

Receiver Operating Characteristic (ROC) Curves

- A test may provide several categories of response rather than simply positive or negative results. In some cases, the results of the test are reported as a continuous variable.
- In either case, the designation of a **cutoff point** for distinguishing a test result as positive versus negative is arbitrary.

TABLE 3.3 Ratings of 109 CT images by a single radiologist vs. true disease status

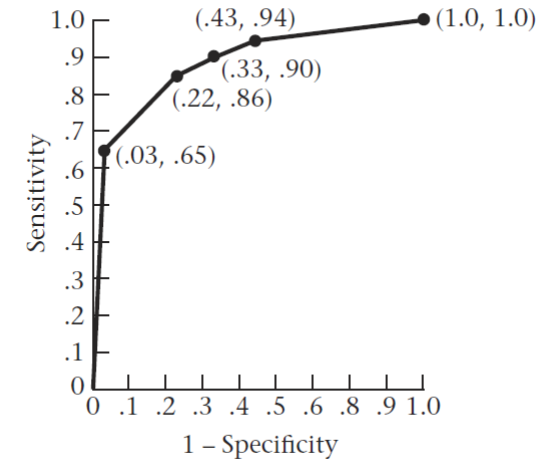
True disease status	CT rating					Total
	Definitely normal (1)	Probably normal (2)	Questionable (3)	Probably abnormal (4)	Definitely abnormal (5)	
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

- This example has no obvious cutoff point to use for designating a subject as positive for disease based on the CT scan.

TABLE 3.4 Sensitivity and specificity of the radiologist's ratings according to different test-positive criteria based on the data in Table 3.3

Test-positive criteria	Sensitivity	Specificity
1 +	1.0	0
2 +	.94	.57
3 +	.90	.67
4 +	.86	.78
5 +	.65	.97
6 +	0	1.0

FIGURE 3.7 ROC curve for the data in Table 3.4*



*Each point represents (1 - specificity, sensitivity) for different test-positive criteria.

- To display above data, we construct a **receiver operating characteristic (ROC) curve**, which is a plot of the **sensitivity versus (1 - specificity)** of a screening test, where the different points on the curve correspond to different **cutoff points** used to designate test-positive.
- **AUC (Area Under the Curve) ≈ 0.875**

Youden Index

- How to Select the Optimal Cutoff Point?
- One common approach is to use the **Youden Index (J)**:
$$J = Sensitivity + Specificity - 1$$

- Alternatively, it can be expressed as:

$$J = TPR - FPR$$

- Based on the data in this example:

Test-Positive Criterion	Youden Index
Rating ≥ 5	0.613
Rating ≥ 4	0.639
Rating ≥ 3	0.574
Rating ≥ 2	0.510

- According to the **maximum Youden Index criterion**, the optimal cutoff point in this example is **Rating ≥ 4** .

Prevalence and Incidence

- **Prevalence** of a disease is the probability of currently having the disease regardless of the duration of time one has had the disease. Prevalence is obtained by dividing the number of people who currently have the disease by the number of people in the study population.
- **Cumulative incidence**, also referred to as incidence, of a disease is the probability that a person with no prior disease will develop a new case of the disease over some specified time period.
- **Incidence rate** of a disease is the rate at which new cases occur per unit of person-time.

Measure	Numerator	Denominator
Point Prevalence	Existing cases	Total population
Cumulative Incidence	New cases	Population initially at risk
Incidence Rate	New cases	Total person-time at risk

Summary

- **Probability Laws:** Addition and multiplication rules are used to calculate probabilities of events.
- **Independence:** Two events are independent if the occurrence of one does not change the probability of the other.
- **Conditional Probability and Relative Risk:** Conditional probability describes probabilities given specific conditions, while relative risk measures the association between exposure and outcome.
- **Diagnostic Test Performance: Sensitivity, specificity, PV^+ (PPV), and PV^- (NPV)** are key measures for evaluating screening tests.
- **ROC Curves:** ROC curves illustrate the trade-off between sensitivity and false positive rate across different classification thresholds.
- **Bayes' Rule:** Bayes' rule combines disease prevalence, sensitivity, and specificity to calculate predictive values.
- **Bayesian Inference:** Prior beliefs are updated using observed data to obtain posterior probabilities.

Key Probability Formulas

Concept	Formula
Addition Law	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Multiplication Law	$P(A \cap B) = P(A)P(B A)$
Independence	$P(A \cap B) = P(A)P(B)$
Conditional Probability	$P(B A) = \frac{P(A \cap B)}{P(A)}$
Total-Probability Rule	$Pr(B) = Pr(B A) \times Pr(A) + Pr(B \bar{A}) \times Pr(\bar{A})$
Relative Risk	$RR = \frac{P(B A)}{P(B \bar{A})}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Bayes' Rule	$P(B A) = \frac{P(A B)P(B)}{P(A)}$