Probability

Overview

- Definitions of Probability
- Sample Space, Events
- Basic Properties
- Joint, Marginal, Conditional Probability
- Rules of Probability
- Screening – Application of Bayes' Rule

NOTHING IN LIFE IS CERTAIN. IN EVERYTHING WE DO, WE GAUGE THE CHANCES OF SUCCESSFUL OUTCOMES, FROM BUSINESS TO MEDICINE TO THE WEATHER. BUT FOR MOST OF HUMAN HISTORY, ***PROBABILITY,*** THE FORMAL STUDY OF THE LAWS OF CHANCE, WAS USED FOR ONLY ONE THING: ***GAMBLING.***

Liber de ludo aleae ("Book on Games of Chance")
by Gerolamo Cardano. Written 1526 (published 1663).
First systematic treatment of probability.
(Included section on effective cheating methods.)

Probability

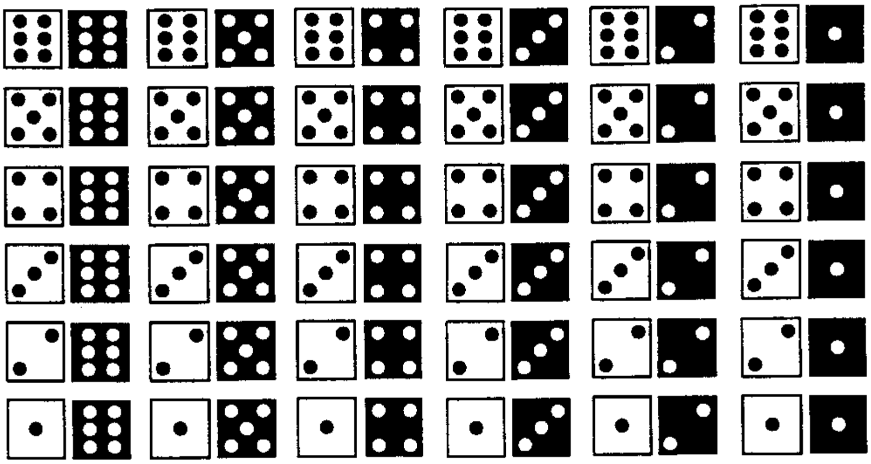
Probability provides a measure of uncertainty associated with the occurrence of events or outcomes

Definitions:

1. **Classical**: P(E) = m/N

   If an event can occur in N mutually exclusive, equally likely ways, and if m of these possess characteristic E, then the probability of E is equal to m/N.

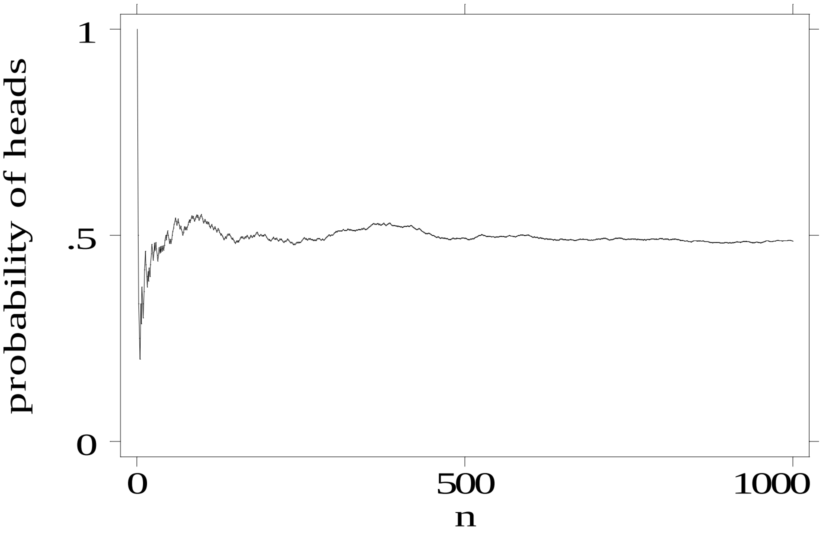
Example: What is the probability of rolling a total of 7 on two dice?



2. **Relative Frequency**: $P(E) \approx m/n$

If a process or an experiment is repeated a large number of times, n, and if the characteristic, E, occurs m times, then the relative frequency, m/n, of E will be approximately equal to the probability of E.

» Around 1900, the English statistician Karl Pearson heroically tossed a coin 24,000 times and recorded 12,012 heads, giving a proportion of 0.5005.



3. **Personal Probability**

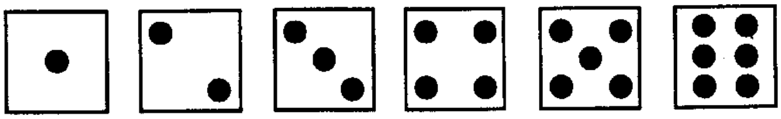
What is the probability of life on Mars?

Sample Space

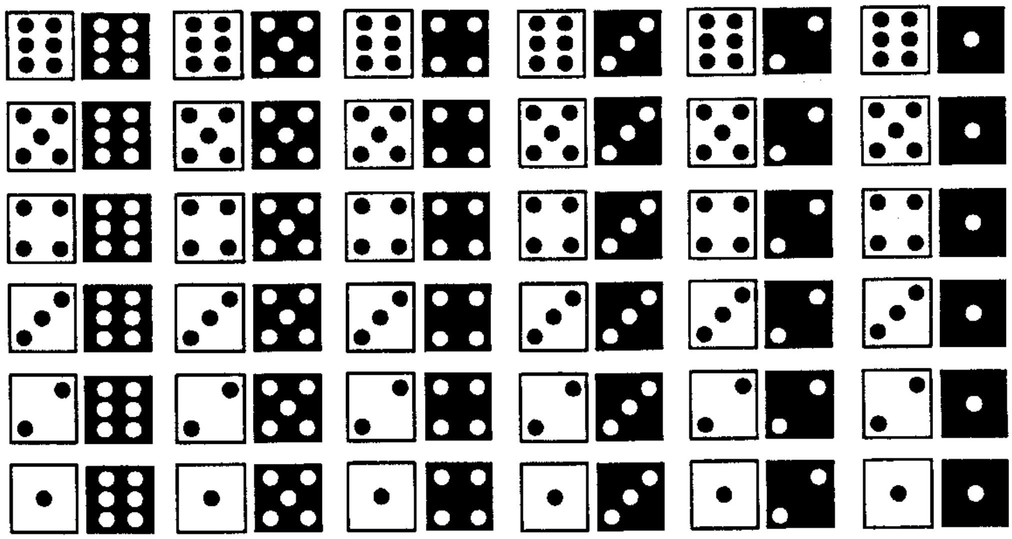
The **sample space** consists of the possible outcomes of an experiment. An **event** is an outcome or set of outcomes.

For a coin flip the sample space is (H,T).

THE SAMPLE SPACE OF THE THROW OF A *SINGLE DIE* IS A LITTLE BIGGER.



AND FOR A *PAIR* OF DICE, THE SAMPLE SPACE LOOKS LIKE THIS (WE MAKE ONE DIE WHITE AND ONE BLACK TO TELL THEM APART):



Sample Space

Key Point # 1:

Whenever you read an article with statistical results, try to identify the sample space. The sample space used by the article may not be the one they want you to think it is.

Example: Woman Wins NJ Lottery Twice

NY Times stated chance was 1 in 17 trillion. True for one particular person purchasing just one ticket each for two different runs.

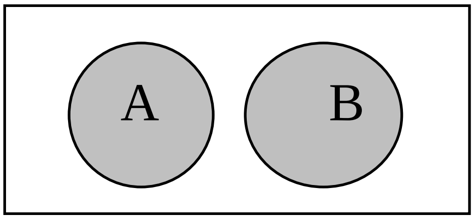
Not true if the question is: "What is the chance that ***someone*** will win the lottery twice in his/her lifetime?" Almost a sure thing!

Back in late 1980's, estimated there is about a 50% chance this will happen in 7-year period.

Diaconis, P., and F. Mosteller. (1989). Methods of Studying Coincidences. *Journal of the American Statistical Association*, **84**(408), 853-861.

Basic Properties of Probability

1. Two events, A and B, are said to be mutually exclusive (disjoint) if only one or the other, but not both, can occur in a particular experiment.



2. Given an experiment with n mutually exclusive events, E_1, E_2, …., E_n, the probability of any event is non-negative and less than 1:

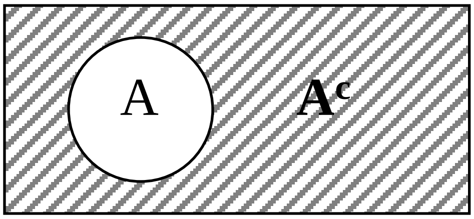
$$0 \leq P(E_i) \leq 1$$

3. The sum of the probabilities of an exhaustive collection (i.e. at least one must occur) of mutually exclusive outcomes is 1:

$$\sum_{i=1}^{n} P(E_i) = P(E_1) + P(E_2) + K + P(E_n) = 1$$

4. The probability of all events other than an event A is denoted by $P(A^c)$ [A^c stands for "A complement"] or $P(\bar{A})$ ["A bar"]. Note that

$P(A^c) = 1 - P(A)$



Basic Properties of Probability

Example: A single die

Consider the following events:

E_1 = roll a 1

E_2 = roll an even number

E_3 = roll a 4, 5 or 6

E_4 = roll a 3 or 5

1) What is $Pr(E_4)$?

2) Are E_2 and E_3 mutually exclusive? E_2 and E_4?

3) Find a mutually exclusive, exhaustive collection of events. Do the probabilities add to 1?

4) What is $Pr(E_4^c)$?

Notation for Joint Probabilities

- If A and B are any two events then we write

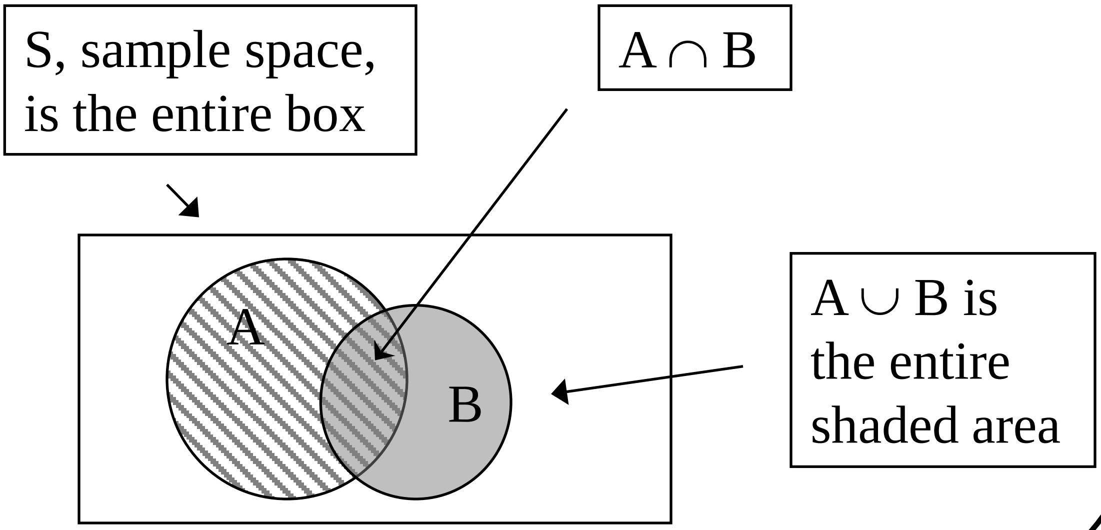
  P(A or B) *or* P(A ∪ B)

  to indicate the probability that event A <u>or</u> event B (or both) occurred.

- If A and B are any two events then we write

  P(A and B) *or* P(AB) *or* P(A ∩ B)

  to indicate the probability that both A <u>and</u> B occurred.



S, sample space, is the entire box

A ∩ B

A ∪ B is the entire shaded area

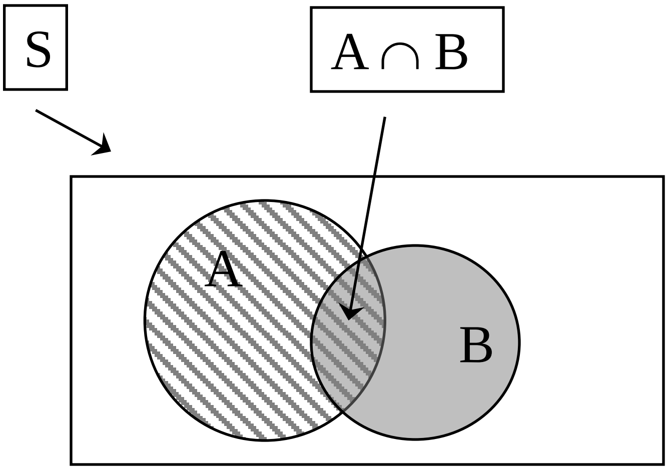
Notation for Joint Probabilities

- If A and B are any two events then we write

  P(A given B) *or* P(A|B)

  to indicate the probability of A among the subset of cases in which B is known to have occurred.

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



Conditional Probability

The conditional probability of an event A given B (i.e. given that B has occurred) is denoted P(A | B).

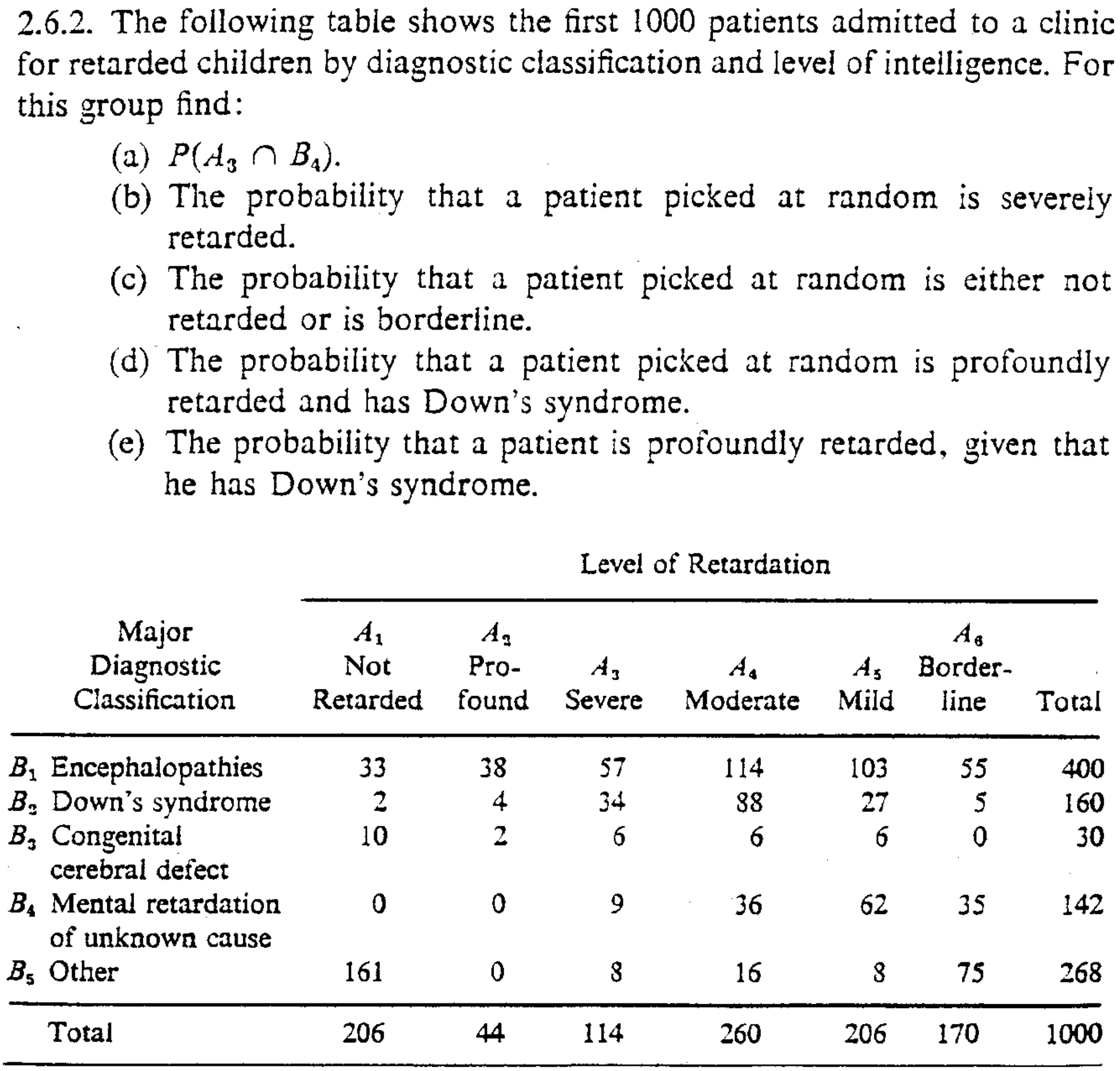
		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	9	80	89
	Neg.	1	9910	9,911
		10	9990	10,000

What is P(test positive)?

What is P(test positive | disease positive)?

What is P(disease positive | test positive)?

Example - Joint Probabilities



2.6.2. The following table shows the first 1000 patients admitted to a clinic for retarded children by diagnostic classification and level of intelligence. For this group find:

(a) $P(A_3 \cap B_4)$.

(b) The probability that a patient picked at random is severely retarded.

(c) The probability that a patient picked at random is either not retarded or is borderline.

(d) The probability that a patient picked at random is profoundly retarded and has Down's syndrome.

(e) The probability that a patient is profoundly retarded, given that he has Down's syndrome.

	Level of Retardation						
Major Diagnostic Classification	A_1 Not Retarded	A_2 Pro-found	A_3 Severe	A_4 Moderate	A_5 Mild	A_6 Border-line	Total
B_1 Encephalopathies	33	38	57	114	103	55	400
B_2 Down's syndrome	2	4	34	88	27	5	160
B_3 Congenital cerebral defect	10	2	6	6	6	0	30
B_4 Mental retardation of unknown cause	0	0	9	36	62	35	142
B_5 Other	161	0	8	16	8	75	268
Total	206	44	114	260	206	170	1000

Joint Probabilities

Key Point # 2:

A probability depends on your definition of the sample space.

The sample space changes with knowledge of the circumstances or what has occurred.

Example:

Car insurance companies don’t set rates based on using probabilities based on ALL drivers, they use probabilities based on categories of drivers
(e.g., Male <22, Female 40-49, etc.)

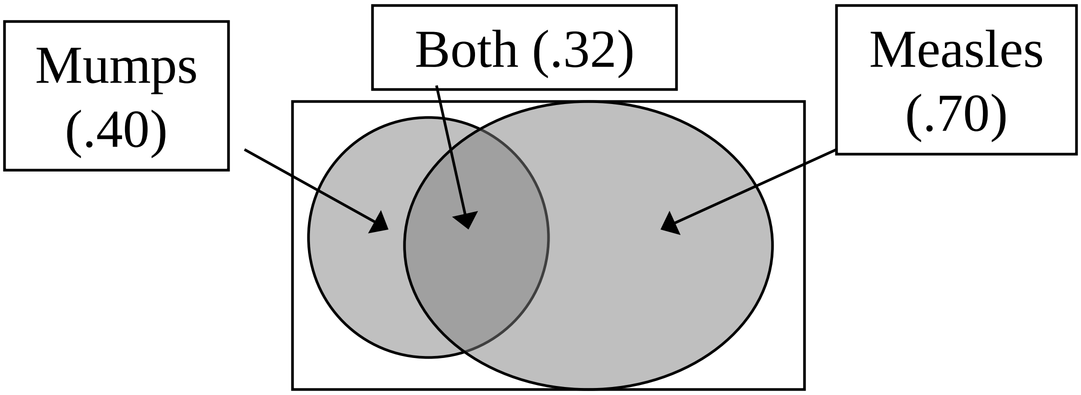
General Probability Rules

- Addition rule

If two events A and B are not mutually exclusive, then the probability that event A or event B occurs is:

$$P(\text{A or B}) = P(A) + P(B) - P(AB)$$

E.g. Of the students at Anytown High school, 40% have had the mumps, 70% have had measles and 32% have had both. What is the probability that a randomly chosen student has had at least one of the above diseases?



General Probability Rules

- Multiplication rule (special case – independence)

If two events, A and B, are “independent” (probability of one does not depend on whether the other occurred) then

P(AB) = P(A)P(B)

E.g.

Suppose
P(mumps) = 40%

P(measles) = 70%

If independent, then we predict
P(mumps, measles) = .4*.7=.28

Easy to extend for independent events A,B,C,…

P(ABC…) = P(A)P(B)P(C)…

General Probability Rules

Two events A and B are said to be independent if and only if

$$P(A|B) = P(A) \text{ or}$$
$$P(B|A) = P(B) \text{ or}$$
$$P(AB) = P(A)P(B).$$

(Note: If any one holds then all three hold)

E.g.

Suppose

P(mumps) = .4, P(measles) = .7

P(both) = .32.

Are the two events independent?

No, because P(mumps and measles) = .32 while

P(mumps) P(measles) = .28

The notion of independent events is pervasive throughout statistics …

General Probability Rules

- Multiplication rule (general)

  More generally, however, A and B may not be independent. The probability that one event occurs may depend on the other event. This brings us back to conditional probability. The general formula for the probability that both A and B will occur is

$$P(AB) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

E.g.

Suppose

P(mumps) = 40%
P(measles | mumps) = 80%

then

P(both) = .80*.40 = .32

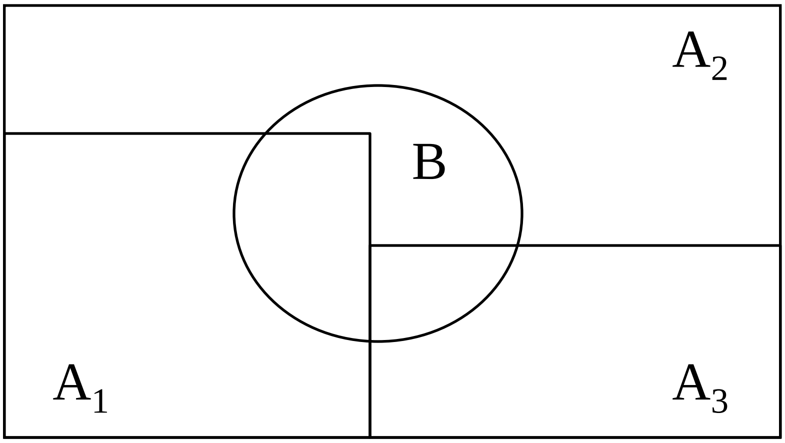
General Probability Rules

- <u>Total Probability Rule</u>

  If $A_1,\ldots A_n$ are mutually exclusive, ***exhaustive*** events, then

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i)$$

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) P(A_i)$$



General Probability Rules

- Total Probability Rule

Example

The following table gives the estimated proportion of individuals with Alzhiemer’s disease by age group. It also gives the proportion of the general population that are expected to fall in the age group in 2030. What proportion of the population in 2030 will have Alzhiemer’s disease?

		Proportion population	Proportion with AD	Hypoth. population	Number affected
Age group	< 65	.80	.00	80,000	0
	65 – 75	.11	.03	11,000	330
	75 – 85	.07	.11	7,000	770
	> 85	.02	.30	2,000	600
				100,000	1700

$$P(AD) = 0*.8 + .03*.11 + .11*.07 + .30*.02 = .017$$

Bayes' Rule

Bayes' rule combines multiplication rule with total probability rule

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

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

$$= \frac{P(B | A_j)P(A_j)}{\sum_{i=1}^{n} P(B | A_i)P(A_i)}$$

We will only apply this to the situation where A and B have two levels each, say, A and $\overline{A}$, B and $\overline{B}$. The formula becomes

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

Screening - An Application of Bayes' Rule

Suppose we have a random sample of a population...

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	30	120
	Neg.	10	970	980
		100	1000	1100

A = disease pos.
B = test pos.

Prevalence = $P(A) = 100/1100 = .091$

Sensitivity = $P(B \mid A) = 90/100 = .9$

Specificity = $P(\overline{B} \mid \overline{A}) = 970/1000 = .97$

PVP = $P(A \mid B) = 90/120 = .75$

PVN = $P(\overline{A} \mid \overline{B}) = 970/980 = .99$

Screening - An Application of Bayes’ Rule

Now suppose we have taken a sample of 100 disease positive and 100 disease negative individuals (e.g. case-control design)

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	3	93
	Neg.	10	97	107
		100	100	200

A = disease pos.
B = test pos.

Prevalence = ???? (not .5!)

Sensitivity = $P(B \mid A) = 90/100 = .9$

Specificity = $P(\overline{B} \mid \overline{A}) = 97/100 = .97$

PVP = $P(A \mid B) = 90/93$ **NO!**

PVN = $P(\overline{A} \mid \overline{B}) = 97/107$ **NO!**

Screening - An Application of Bayes Rule

A = disease pos.
B = test pos.

Assume we know, from external sources, that P(A) = 100/1100. Then for every 100 disease positives we should have 1000 disease negatives …. 1:10.

Make a mock table …

		Disease Status		
		Pos.	Neg.	
Test Result	Pos.	90	3 × 10	120
	Neg.	10	97 × 10	980
		100	100 × 10	1100

$$\text{PVP} = \frac{90}{90+3\times10} = .75$$

Screening - an application of Bayes Rule

Now, use Bayes rule …

$$
\begin{aligned}
\text{PVP} = \text{P(A|B)} &= \frac{\text{P(B|A)P(A)}}{\text{P(B)}} \\
&= \frac{\text{P(B|A)P(A)}}{\text{P(B|A)P(A)} + \text{P(B|}\overline{\text{A}}\text{)P(}\overline{\text{A}}\text{)}} \\
&= \frac{.9 \times {}^{100}\!/_{1100}}{.9 \times {}^{100}\!/_{1100} + .03 \times {}^{1000}\!/_{1100}} \\
&= \frac{.9 \times 100}{.9 \times 100 + .03 \times 1000} = .75
\end{aligned}
$$

Summary

- Probability - meaning
  - 1) classical
  - 2) frequentist
  - 3) subjective (personal)
- Sample space, events
- Mutually exclusive, independence
- and, or, complement
- Joint, marginal, conditional probability
- Probability - rules
  - 1) Addition
  - 2) Multiplication
  - 3) Total probability
  - 4) Bayes
- Screening
  - sensitivity
  - specificity
  - predictive values

Problems

1. If allelle A has frequency 3/4 and allelle a has frequency 1/4 , what are the prevalences of the 3 genotypes AA, Aa and aa in the population (assuming random mating)?
2. A certain operation has a survival rate of 70%. If this operation is performed independently on three different patients, what is the probability all three operations will fail?
3. Suppose an influenza epidemic strikes a city. In 10% of (two parent) families the mother has influenza (event A); in 10% of families the father has influenza (event B) and in 1% of families both the mother and father have influenza.
   a) Are the events A and B independent?
   b) What is the probability neither the mother nor father have influenza?
4. The following table gives the probability of disease for different alleles of a gene (penetrances). What is the predcited probability of disease on a randomly selected individual if you have no genetic information? (Hint: use the total probability rule)

Allele	Proportion with this allele	Probability of disease with this allele
A1	.0004	.540
A2	.0059	.813
A3	.0855	.379
A4	.9082	0.0

5. In a group of symptomatic women attending a clinic, some had cervical infections with *Chlamydia trachomatis* (C) or *Neisseria gonorrhea* (G), and some were harboring both organisms. Seven women had C only, 5 women had G only and 8 women had both (B).
   a) What is the probability of any chlamydia (C) present?
   b) What is the probability of any gonorrhea (G) present?
   c) What is the probability of any gonorrhea (G) or chlamydia (C) present?
   d) Are gonorrhea and chlamydia mutually exclusive?

Problems

6) The following table summarizes a famous study by Jerushalmy et al that sparked controversy concerning the value of various screening procedures for disease detection.

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

a) If one of the 1820 records were randomly selected, what is the probability it would be a person with TB?

b) For a randomly selected record, what is the probability that it belongs to a person who has TB and has a positive X-ray?

c) If you are told that a randomly selected record is for a person with a positive X-ray, what is the probability that it belongs to a person with TB?

d) What is the probability that a randomly selected record belongs to a person with TB or a person with a positive X-ray?

7) Estimates of the proportion of individuals with Alzheimer's disease (AD) in various age and gender groups is given in the following table. Suppose an unrelated 77 year old man, 76 year old woman and 82 year old woman are selected from the community represented in this sample. Each will be tested for AD.

Age group	Males	Females
65-69	0.016	0.0
70-74	0.0	0.022
75-79	0.049	0.023
80-84	0.086	0.078
85+	0.35	0.279

a)The sample space for this "experiment" consists of all possible outcomes of the testing. List these (hint: there are 8 possible outcomes).

b)What is the probability all three have AD?

c)What is the probability at least one has AD?

d)What is the probability exactly one has AD?

Solutions

1) P(AA) = (3/4)*(3/4)= 9/16   P(Aa) = 2*3/16 = 6/16   P(aa) = 1/16

2) P(fail,fail,fail) = P(fail)P(fail)P(fail) = .3*.3*.3 = .027

3) a) Yes, since .1*.1 = .01

   b) P(neither) = .9*.9 = .81

4) Prob = .0004*.54 + .0059*.813 + .0855*.379 + .9082*0 = .037

5) a) (7+8)/20 = .75

   b) (5+8)/20 = .65

   c) 20/20 = 1

   d) No, can have both

6) a) 30/1820

   b) 22/1820

   c) 22/73

   d) (30+73-22)/1820

7) a) Let A = has AD; a = does not have AD

77yo	76yo	82yo	Prob
A	A	A	.049*.023*.078
A	A	a	.049*.023*(1-.078)
A	a	A	etc
A	a	a	
a	A	A	
a	A	a	
a	a	A	
a	a	a	

   b) .049*.023*.078 = 8.7906e-05

   c) 1 – P(aaa) = 1 – (1-.049)(1-.023)(1-.078) = .143

   d) P(Aaa)+P(aAa)+P(aaA) = .136