Statistical Theory

- Basic Probability Concepts
- Probability Distributions
- Population and Samples

Basic Probability Concepts

Probability

- Chance of observing a particular outcome
- Likelihood of an event

Assumes a "stochastic" process: i.e.. the outcome is not predetermined - there is an element of chance

How are probabilities assigned?

- Classical (based on gambling ideas)
- Relative Frequency (in the long run. . .)
- Personal (personal assessment)

Examples

- Classical

Rolling a die -

There are 6 possible outcomes:

S={1, 2, 3, 4, 5, 6}. Each is equally likely

Pr(i) = 1/6, i=1,2,...,6.

- **Relative Frequency**

Predicting the weather – "Past records show that 70 times out of 100 when the wind and temperature patterns were like this it rained. Therefore, there is a 70% chance of rain today."

Recent experience tells us that 80 of the last 100 patients having surgery had no post-operative complications. The next patient is told there is a 20% chance of a complication related to the surgery.

- Personal

Personal assessment of which sports team will win a match.

Personal assessment of which is more likely to provide a cure – traditional healing or modern medicine.

9/1/02 7

- **An Objectivist uses the classical and/or relative frequency definition to determine probabilities.**
- **A Subjectivist uses all three methods to assess the likelihood of an event.**

9/1/02 8

Example: Coin toss

- Fair coin, fairly tossed. What is the probability of the coin landing "head" up?
- Pr(H)

9/1/02 9

Classical

- Assume equally likely outcomes (classical)

  Pr(H) = 0.5

9/1/02 10

Relative frequency

- In the previous 1,000 tosses of this coin, heads have turned up 542 times.
- Pr(H) = 0.542

9/1/02 11

Personal

- This is my lucky coin; it usually comes up heads.
- Pr(H) = unspecified, but maybe 0.75

9/1/02 12

- **A <u>Probability Model</u> is the set of assumptions used to assign probabilities to each outcome in a set of possible outcomes.**
- **A <u>Probability Distribution</u> defines the relationship between the outcomes and their probability of occurrence.**

Characteristics of a probability distribution

- All possible events included: events are mutually exclusive and collectively exhaustive
- Probability of one event is between 0 and 1, inclusive
- Sum of probabilities of all events is 1.

Notation

- Let e_i denote one event
- S = $\{e_i\}$ i=1,2,...s, represents the entire set of events.
- S = $\{e_1, e_2, \ldots, e_s\}$
- $0 \leq pr(e_i) \leq 1$
- $pr(e_1) + pr(e_2) + \ldots + pr(e_s) = 1$

Example: Sex of a baby

e_1=girl
e_2=boy
S={e_1,e_2}
Pr(e_1)=1/2
Pr(e_2)=1/2
Pr(e_1)+Pr(e_2)=1

Example: Sexes of twin pairs

e_1 = [boy, boy]
e_2 = [boy, girl]
e_3 = [girl, boy]
e_4 = [girl, girl]
S = {e_1 , e_2 , e_3 , e_4}

exactly 1 girl = {e_2, e_3}
at least 1 girl = {e_2, e_3, e_4}
same sex = {e_1, e_4}
2 girls = {e_4}

Assume fraternal twins

$pr(e_1) = 1/4$
$pr(e_2) = 1/4$
$pr(e_3) = 1/4$
$pr(e_4) = 1/4$
Sum 1

$pr\ (e_2,e_3) = 1/2$
$pr\ (e_2, e_3, e_4) = 3/4$
$pr\ (e_1, e_4) = 1/2$
$pr\ (e_4) = 1/4$

9/1/02 19

Assume identical twins

$pr(e_1) = 1/2$
$pr(e_2) = 0$
$pr(e_3) = 0$
$pr(e_4) = 1/2$
Sum 1

$pr\ (e_2,e_3) = 0$
$pr\ (e_2,e_3 , e_4) = 1/2$
$pr\ (e_1 , e_4) = 1$
$pr\ (e_4) = 1/2$

9/1/02 20

Some definitions, notation and rules:

- Marginal probability:
  Probability of a single event
  Probability a person's systolic BP < 140

  Pr (BP< 140)

9/1/02 21

- Conditional probability
  - Probability of an event conditional on the occurrence (or non-occurrence) of another event
  - Probability that a man's systolic BP < 140.
  - Given that he is a man, probability that systolic BP < 140.
  - Pr(BP< 140| male)

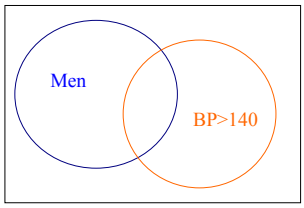
- Joint probability
  - Probability of the joint occurrence of two events
  - Probability that a person has systolic BP < 140 and is a man
  - Pr (BP < 140, male)

$$\Pr(BP<140) = \frac{\#\text{ people with } BP<140}{\#\text{ people}}$$

$$\Pr(BP<140, \text{male}) = \frac{\#\text{ males with } BP<140}{\#\text{ people}}$$

$$\Pr(BP<140|\text{male}) = \frac{\#\text{ males with } BP<140}{\#\text{ males}}$$

Venn Diagram



Multiplication Rule

For two events A and B,

$Pr(A,B) = Pr(A|B)*Pr(B)$

$Pr(A,B) = Pr(B|A)*Pr(A)$

Other relationships that follow

- $Pr(A|B) = \frac{Pr(A,B)}{Pr(B)}$
- $Pr(B|A) = \frac{Pr(A,B)}{Pr(A)}$

- Independence:
  Two events are said to be independent if the occurrence of one has no effect on the occurrence of the other

  If Pr(A|B) = Pr(A) then A and B are independent

Addition rule

- For two events A and B,

Pr(A or B) = Pr(A) + Pr(B) - Pr(A,B)

If the events are mutually exclusive, then
Pr(A,B) = 0 and
Pr(A or B) = Pr(A) + Pr(B)

If two events are independent then
Pr(A, B) = Pr(A)*Pr(B)

Proof

Pr(A, B) = Pr(A|B)*Pr(B)

If A and B are independent then
Pr(A|B) = Pr(A)

and Pr(A,B) = Pr(A)*Pr(B)

Example: Job strain and Gender

	Job strain	No job strain	Total
Men	305	841	1146
Women	183	362	545
Total	488	1203	1691

$$\Pr(M) = \frac{1146}{1691} = 0.6777$$

$$\Pr(M|JS) = \frac{305}{488} = 0.625$$

$$\Pr(M,JS) = \frac{305}{1691} = 0.1804$$

Multiplication rule

$\Pr(M,JS) = \Pr(M|JS)*\Pr(JS)$

$\Pr(M|JS) = 0.625$
$\Pr(JS) = 0.2886$
$\Pr(M|JS)*\Pr(JS) = 0.1804$

Addition Rule

Pr(M or JS) = Pr(M) + Pr(JS) - Pr(M,JS)

Pr(M or JS) = $\frac{1329}{1691}$ = 0.7859

Pr(M) + Pr(JS) - Pr(M,JS) = 0.6777 + 0.2886 - 0.1804
= 0.7859

9/1/02 34

Using the multiplication and addition rules, we can often obtain probabilities through algebraic manipulation of other probabilities.

9/1/02 35

Bayes's Theorem

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A)+Pr(B|A^{-})Pr(A^{-})}$$

9/1/02 36

Multiplication rule: $Pr(A|B) = \frac{Pr(A,B)}{Pr(B)}$

Addition rule: $Pr(B) = Pr(B,A) + Pr(B,A^-)$

$$Pr(A|B) = \frac{Pr(A,B)}{Pr(A,B)+Pr(A^-,B)}$$

Multiplication rule (again): $Pr(A,B) = Pr(B|A)Pr(A)$

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A)+Pr(B|A^-)Pr(A^-)}$$

Uses of probability

- Classical probability models used in statistical theory
- Frequency probability models used in epidemiology and other applications
  - Sensitivity, Specificity, Predictive value
  - Relative risk
  - Interpretation of other rates

Application: Diagnostic Testing

- Validity of a screening test
  - Sensitivity
  - Specificity
- Predictive value of a screening test

9/1/02 40

Validity

- Sensitivity: the ability of the test to correctly identify those who **have** the disease
- Specificity: the ability of the test to correctly identify those who **do not have** the disease

9/1/02 41

- Sensitivity = $Pr(T^+|D^+)$
- Specificity = $Pr(T^-|D^-)$

9/1/02 42

	Disease (D+)	No disease (D-)	Total
Test +	a	b	a+b
Test -	c	d	c+d
	a+c	b+d	N=a+b+c+d

9/1/02 43

- Sensitivity = Pr(T+|D+) = $\frac{a}{a+c}$
- Specificity = Pr(T-|D-) = $\frac{d}{b+d}$

9/1/02 44

Example: Screening for Prostate Cancer

	Ca	No Ca	
PSA +	57	1,712	1,769
PSA -	21	13,126	13,147
	78	14,838	14,916

9/1/02 45

- Sensitivity $= \frac{57}{78} = 0.73$
- Specificity $= \frac{13126}{14838} = 0.88$

Predictive value

- Condition on test results
- Useful for physician planning, etc.
  - Of the patients who test positive, what percentage actually have the disease
- Measures the practical utility of implementing a test in a large setting.

- Predictive value of a positive test: PV+ = Pr(D+|T+)
- Predictive value of a negative test: PV- = Pr(D-|T-)

Prostate cancer

- PV+ $= \frac{57}{1769} = 0.032$
- PV- $= \frac{13126}{13147} = 0.998$

9/1/02 49

For rare diseases, we often evaluate a test on a sample of cases and and equal size sample of non-cases (controls).

	Disease	No disease	
Test +	a	b	a+b
Test -	c	d	c+d
	n	n	N=2n

9/1/02 50

- Sensitivity = a/n
- Specificity = b/n

But, cannot estimate PV+ or PV- because Pr(D+) artificially set to 0.5

9/1/02 51

Use Bayes's theorem

- If disease prevalence is known, PV+ can be computed from sensitivity, specificity and disease prevalence.

$$Pr(D+|T+)=\frac{Pr(T+|D+)Pr(D+)}{Pr(T+|D+)Pr(D+)+Pr(T+|D-)Pr(D-)}$$

9/1/02 52

HIV Example

- The sensitivity and specificity of the HIV screening test are very high.
- Should the general public be screened for HIV?
- What is the predictive value of the HIV test?

9/1/02 53

	HIV +	HIV -	Total
Test +	99	4	103
Test -	1	96	97
	100	100	200

9/1/02 54

- Sensitivity = 0.99
- Specificity = 0.96
- Disease prevalence is about 2%

9/1/02 55

Predictive value by Bayes's Theorem

$$PV+ = \Pr(D+ \mid T+)$$
$$= \frac{\Pr(T+ \mid D+)\Pr(D+)}{\Pr(T+ \mid D+)\Pr(D+) + \Pr(T+ \mid D-)\Pr(D-)}$$
$$= \frac{0.99 * 0.02}{0.99 * 0.02 + (1-0.96) * (1-0.02)}$$
$$= \frac{0.0198}{0.0198 + 0.0392} = 0.34$$

9/1/02 56