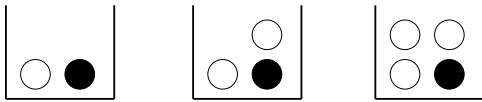


Statistics and Probability

Probability Basics and Bayes' Rule



Box 1

Box 2

Box 3

<http://www.isrec.isb-sib.ch/~darlene/geneve/>

27 March 2007

Statistics and Probability

Lec 2

Outline

- More aspects of the R language
- Probability rules
- Equally likely outcomes
- Counting rules
- Conditional probability
- Bayes' rule

27 March 2007

Statistics and Probability

Lec 2

R: variable types

```
> a <- 49                                     numeric
> sqrt(a)
[1] 7

> a <- "The dog ate my homework"             character
> sub("dog","cat",a)                         string
[1] "The cat ate my homework"

> a <- (1+1==3)
> a                                           logical
[1] FALSE
```

27 March 2007

Statistics and Probability

Lec 2

R: factors

- Categorical variables in R should be specified as *factors*
- Factors can take on a limited number of values, called *levels*
- Levels of a factor may have a natural order
- Functions in R for creating factors: `factor()`, `ordered()`

27 March 2007

Statistics and Probability

Lec 2

R: missing values

- Variables of each data type can also take the value **NA** (for **N**ot **A**vailable)
 - NA is not the same as 0
 - NA is not the same as "" (blank, or empty string)
 - NA is not the same as FALSE
- Computations involving NA:

```
> 1+NA
[1] NA
> max(c(NA, 4, 7))
[1] NA
> max(c(NA, 4, 7), na.rm=TRUE)
[1] 7
```

27 March 2007

Statistics and Probability

Lec 2

R: NA in statistical functions

- Different functions have different defaults on handling NA values
- For single vector functions (e.g. `mean`, `var`, `sd`), give the *argument* `na.rm=TRUE`
- Read the help documentation for each function that you use!

27 March 2007

Statistics and Probability

Lec 2

R: functions and operators

- Functions 'do things' with data
 - Input: function *arguments* (0,1,2,...)
 - Output: function *result* (exactly one)
- Exceptions:
 - Functions may also use data that sits in other places, not just in the argument list
 - Functions may also do things other than return a result ("*side effects*")

27 March 2007

Statistics and Probability

Lec 2

R: logical and relational operators

- `==` Equal to
- `!=` Not equal to
- `<` Less than
- `>` Greater than
- `<=` Less than or equal to
- `>=` Greater than or equal to
- `is.na(x)` Missing?
- `&` Logical AND
- `|` Logical OR
- `!` Logical NOT

27 March 2007

Statistics and Probability

Lec 2

R: vectors

- *vector*: an ordered collection of data of the *same type*:

```
> a <- c(7,5,1)
> a*2
[1] 14 10 2
```
- *Example*: the mean spot intensities of all 15488 spots on a chip; in R, a vector of 15488 numbers
- A single number is the special case of a vector with 1 element
- Other vector types: character strings, logical

27 March 2007

Statistics and Probability

Lec 2

R: matrices and arrays

- *matrix*: a rectangular table of data of the *same type*
- *Example*: expression values of 10000 genes for 30 tissue biopsies; in R, a matrix with 10000 *rows* and 30 *columns*
- *array*: a multiply indexed collection of data entries (e.g. a 'stack' of matrices)
- *Example*: the red and green foreground and background values for 20000 spots on 120 chips; in R, a 4 x 120 x 120 (3D) array

27 March 2007

Statistics and Probability

Lec 2

R: lists

- *vector*: an ordered collection of data of the *same type*; individual elements accessed with `[]`

```
> a <- c(7,5,1)
> a[2]
[1] 5
```
- *list*: an ordered collection of data of *arbitrary types*; individual elements accessed with `$`

```
> doe <-
  list(name="john",age=28,married=FALSE)
> doe$name
[1] "john"
> doe$age
[1] 28
```

27 March 2007

Statistics and Probability

Lec 2

R: data frames

- *data frame*: the type of R object normally used to store a data set
- A data frame is a rectangular table with rows and columns
 - data *within* each column has the *same type* (e.g. number, character, logical)
 - different columns may have *different types*
- *Example*:

```
> tumor.info
      localisation tumorsize progress
XX348    proximal      6.3    FALSE
XX234    distal       8.0     TRUE
XX987    proximal     10.0    FALSE
```

27 March 2007

Statistics and Probability

Lec 2

R: making data frames

- Data frames can be created in R by *importing* a data set
- A data frame can also be created from pre-existing variables
- Example:**

```
> localisation<-c("proximal","distal","proximal")
> tumorsize<- c(6.3,8,10)
> progress<-c(FALSE,TRUE,FALSE)
> tumor.info<-
  data.frame(localization,tumorsize,progress)
> rownames(tumor.info)<-c("XX348","XX234","XX987")
> tumor.info$tumorsize
[1] 6.3 8.0 10.0
```

27 March 2007

Statistics and Probability

Lec 2

R: subsetting (or indexing)

- A very powerful feature of R is the *subset operator* []
- Allows access to *individual elements* of a vector, matrix, array or data frame are by specifying their index(es), or their name(s)
- Examples:**

```
> tumor.info[3, 2]
[1] 10
> tumor.info["XX987", "tumorsize"]
[1] 10
> tumor.info["XX987",]
  localisation tumorsize progress
XX987 proximal 10 FALSE
```

27 March 2007

Statistics and Probability

Lec 2

R: more on subsetting

```
> tumor.info[c(1,3),]
  localisation tumorsize progress
XX348 proximal 6.3 FALSE
XX987 proximal 10.0 FALSE
subset rows by a vector of indices

> tumor.info[c(TRUE,FALSE,TRUE),]
  localisation tumorsize progress
XX348 proximal 6.3 FALSE
XX987 proximal 10.0 FALSE
subset rows by a logical vector

> tumor.info$localisation
[1] "proximal" "distal" "proximal"
subset a column

> tumor.info$localisation=="proximal"
[1] TRUE FALSE TRUE
comparison resulting in logical vector

> tumor.info[ tumor.info$localisation=="proximal", ]
  localisation tumorsize progress
XX348 proximal 6.3 FALSE
XX987 proximal 10.0 FALSE
subset the selected rows
```

27 March 2007

Statistics and Probability

Lec 2

R: importing and exporting data

- Many ways* to get data into and out of R
- One straightforward way is to use *tab-delimited text files* (e.g. save an Excel sheet as tab-delimited text, for easy import into R)
- Useful R functions: `read.delim()`, `read.table()`, `read.csv()`, `write.table()`
- Example:**

```
> x <- read.delim("filename.txt")
> write.table(x, file="x.txt",
  sep="\t")
```

27 March 2007

Statistics and Probability

Lec 2

R: light modeling introduction

- Can read `~` as 'described (or modeled) by'
- Example:** to make separate boxplots based on a grouping variable

```
> boxplot(my.response ~ group.variable)
```

27 March 2007

Statistics and Probability

Lec 2

Sample space

- '*Experiment*' whose *outcome* cannot be predicted with certainty in advance
- May, however, know set of all *possible* outcomes, the *sample space* S
- Simple examples:
 - One toss of a coin: $S = \{H, T\}$
 - Two tosses of a coin: $S = \{(H,H), (H,T), (T,H), (T,T)\}$
 - One toss of a die: $S = \{1, 2, 3, 4, 5, 6\}$
- S can be *discrete* or *continuous*

27 March 2007

Statistics and Probability

Lec 2

Events

- Informal: the 'event' that a tossed coin lands Heads; sum of the numbers on two dice is 7, *etc.*
- Formal definition: an *event* is any subset of the sample space S
- Normally use capital letters to denote events: A, B , *etc.*

27 March 2007

Statistics and Probability

Lec 2

Probability

- The *long run frequency interpretation* of *chance* or *probability* says that the chance of an event is the percentage (or proportion) of the time we expect the event to occur
- This is the most commonly used definition of probability, but is not the only one

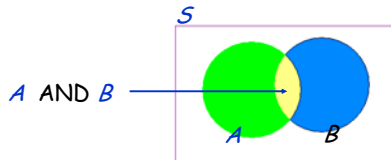
27 March 2007

Statistics and Probability

Lec 2

Venn diagrams

- A *Venn diagram* is a picture used in set theory (which probability theory makes use of)
- Can think of probabilities as relative areas of a Venn diagram



27 March 2007

Statistics and Probability

Lec 2

Probability rules

- $0 \leq P(A) \leq 1$
- $P(S) = 1$
- *Mutually exclusive (ME)* events cannot all occur together in the same trial
 - *Example*: toss 1 die; A = number is even; B = number is 5
- *Addition Rule*: If A, B, C, \dots are *mutually exclusive*, or *incompatible*, events, then

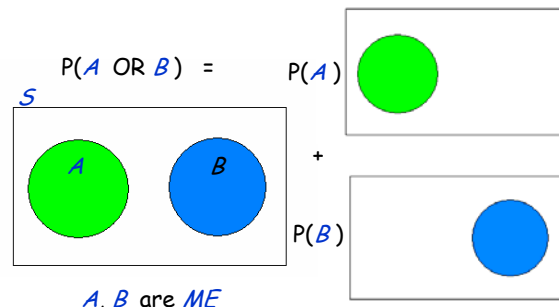
$$P(A \text{ OR } B \text{ OR } C \text{ OR } \dots) = P(A) + P(B) + P(C) + \dots$$

27 March 2007

Statistics and Probability

Lec 2

Illustration for addition rule

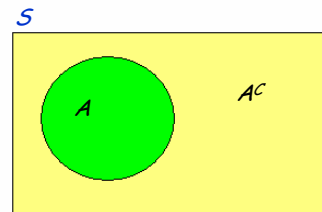


27 March 2007

Statistics and Probability

Lec 2

Complement



27 March 2007

Statistics and Probability

Lec 2

A few more rules

- If an event A is *certain*, then $P(A) = 1$; if A is *impossible*, then $P(A) = 0$
- The *complement* of an event A is its 'opposite': everything in S that is not part of A
- *Complement Rule*: $P(A^c) = 1 - P(A)$
- *General Addition Rule* (also referred to as the *inclusion-exclusion formula*):

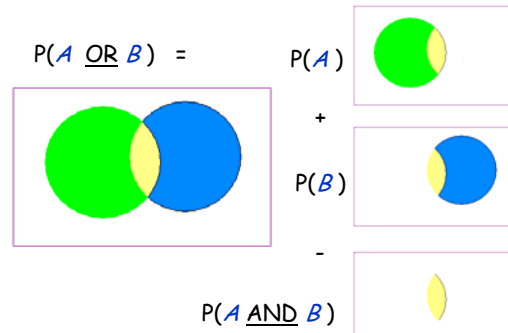
$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$$

27 March 2007

Statistics and Probability

Lec 2

General addition rule.



27 March 2007

Statistics and Probability

Lec 2

Equally likely outcomes

- Simplest case for calculating probabilities is for a *finite number* N of possible outcomes, and each outcome is *equally likely*
- Simple examples:
 - Toss a fair coin one time
 - Toss 2 fair dice one time
- Then the probability of each outcome is $1/N$ (*simple* sample space)
- By the addition rule, the probability of an event A consisting of M outcomes is M/N

27 March 2007

Statistics and Probability

Lec 2

Equally likely outcomes: Example

- Q : If two fair dice are rolled, what is the probability that the sum of the numbers is 7?
- A : If we assume that each of the 36 possible outcomes is equally likely, then the problem is easy to solve
- There are 6 possible outcomes which result in a sum of 7; these are:
 $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$
- So, the probability is $6/36 = 1/6$

27 March 2007

Statistics and Probability

Lec 2

Another simple example

- Q : Three fair coins are simultaneously tossed; find the probability of exactly two heads
- A : Assume that each of the 8 possible outcomes is equally likely (reasonable for fair coins randomly tossed):
 HHH HHT HTH HTT THH THT TTH TTT
- So, $P(2 H) = 3/8$

27 March 2007

Statistics and Probability

Lec 2

Another simple example, cont.

- What if we had regarded the possible outcomes as: 0 H, 1 H, 2 H, 3 H, so that there are 4 possible outcomes?
- Then the sample space is not simple, because these outcomes are *NOT* equally likely
- So, *cannot* say that $P(2 H) = 1/4$

 **WRONG!!!!**

27 March 2007

Statistics and Probability

Lec 2

Why counting rules?

- Recall the definition of probability in the case of equally likely outcomes:

If there are M equally likely outcomes in the event A , and N equally likely outcomes in total, then $P(A) = M/N$

- When the number of outcomes is large, counting rules help to figure out M and N

27 March 2007

Statistics and Probability

Lec 2

Basic principle of counting

- Suppose that two experiments are to be performed. Then if experiment 1 can result in any one of m possible outcomes and if, for each outcome of experiment 1, there are n possible outcomes of experiment 2, then together there are $m*n$ possible outcomes of the two experiments.
- Can generalize to any number of independent experiments

27 March 2007

Statistics and Probability

Lec 2

Example

- Q:** A university planning committee consists of 3 doctoral students, 4 administrators, 5 faculty members, and 2 undergraduates. A subcommittee of 4, consisting of 1 individual from each group, is to be chosen. How many subcommittees are possible?
- A:** From the generalized version of the basic principle, there are $3*4*5*2 = 120$ possible subcommittees

27 March 2007

Statistics and Probability

Lec 2

Some definitions

- A **permutation** is an arrangement of objects in a definite order
- A **combination** is a selection of objects without regard to order
- Does the order matter?
 - Yes → Permutation
 - No → Combination

27 March 2007

Statistics and Probability

Lec 2

Permutations

- Q:** How many different ordered arrangements of the letters a, b, c are possible?
- A:** There are a few ways to figure this out:
 - Direct enumeration:** list all possibilities and count
abc, acb, bac, bca, cab, cba → 6
 - Basic counting rule:** the first can be any of the 3, the second can then be chosen from the other 2, and the third is 'chosen' from the remaining 1
 $3*2*1 = 6$

27 March 2007

Statistics and Probability

Lec 2

Counting permutations

- Suppose there are n objects. The objects can be arranged in $n*(n-1)*(n-2)*...*2*1 = n!$ different orders
- The number $n!$ is pronounced '***n factorial***'
- $0! = 1$
- Follows from reasoning of basic principle

27 March 2007

Statistics and Probability

Lec 2

Counting perms of r out of n objects

- The number of unique arrangements of r objects out of n distinct objects is $n * (n-1) * \dots * (n-r+1) = n!/(n-r)!$
- *Example:*
- *Q:* 8 swimmers are competing for first, second, and third place. In how many ways can the top 3 swimmers finish?
- *A:* $8*7*6 (= 8!/5!) = 336$

27 March 2007

Statistics and Probability

Lec 2

Counting combinations - example

- How many different groups of 3 could be selected from the 5 objects A, B, C, D, E?
- 5 ways to select the first, 4 to then select the second, and 3 ways to select the third, so $5*4*3$ ways to choose when order matters
- But, every group of 3 (e.g. A, B, C) is counted $6 (=3!)$ times: ABC, ACB, BAC, BCA, CAB, CBA
- So, total number of groups when *ignore order* is $5*4*3/(3*2*1) = 10$

27 March 2007

Statistics and Probability

Lec 2

Counting combinations

- In general, the number of different groups of r objects out of n is $n * (n-1) * \dots * (n-r+1)/r! = n!/(r! (n-r)!)$
- This number is written $\binom{n}{r}$, and is pronounced '*n choose r*'
- *Example:* How many 5 card poker hands are possible (out of a deck of 52 cards)?
There are $\binom{52}{5} = 2,598,960$ poker hands

27 March 2007

Statistics and Probability

Lec 2

(BREAK)

27 March 2007

Statistics and Probability

Lec 2

Dice tossing

- Say we toss 2 dice, and each of the 36 possible outcomes is equally likely
- What is the chance that the sum is 8?
 - The event {sum = 8} corresponds to the outcomes (2,2)... so has (unconditional) probability = ??

27 March 2007

Statistics and Probability

Lec 2

Sample space

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

27 March 2007

Statistics and Probability

Lec 2

A new question

- Now I tell you that the first die is 3. Given this information, what is $P(\text{sum} = 8)$?
- Recall our basic calculation of probability when outcomes are equally likely: $P = M/N$
- How many outcomes are there in the sample space? Now, we know that the first die is a 3, so the sample space is reduced to outcomes where the first toss is 3:

27 March 2007

Statistics and Probability

Lec 2

New sample space

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

27 March 2007

Statistics and Probability

Lec 2

Solution

- There are $??$ outcomes in this new sample space
- OF the outcomes in the new sample space, how many correspond to the event: $\text{sum} = 8$?

27 March 2007

Statistics and Probability

Lec 2

Solution

(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

27 March 2007

Statistics and Probability

Lec 2

Conditional probability

- Another way of saying this: *Given* that the first die is a 3, the (conditional) probability that the sum is 8 is $1/6$
- Let $A = \text{'sum is 8'}$; $B = \text{'1st is 3'}$
- We write $P(A | B) = 1/6$, and read '|' as 'given'
- Formally,

$$P(A | B) = P(A \text{ AND } B) / P(B)$$
- $P(A | B)$ is *not defined* if $P(B) = 0$

27 March 2007

Statistics and Probability

Lec 2

Why conditional probability

- The concept of *conditional probability* is important for two main reasons:
 - When interest is in calculating probabilities when some *partial information* concerning the result of an experiment is available
 - Useful as a *tool* to more easily compute some other desired probability

27 March 2007

Statistics and Probability

Lec 2

Another example

- A coin is flipped twice. If we assume that all 4 possible outcomes are equally likely, what is the conditional probability that both flips land heads, given that the first one does?

$$P(H,H | (H, ??)) = \frac{P(H,H)}{P(H,H) \text{ OR } (H,T)}$$

$$= (1/4) / (2/4) = 1/2$$

27 March 2007

Statistics and Probability

Lec 2

General multiplication rule

- The *general multiplication rule* says that the chance both events happen is the chance that the first happens, multiplied by the chance that the second happens given that the first has happened

- This can be extended to any number of events:

$$P(A_1 \text{ AND } A_2 \text{ AND } \dots \text{ AND } A_n)$$

$$= P(A_1) * P(A_2 | A_1) * P(A_3 | A_1, A_2) * \dots$$

$$* P(A_n | A_1, A_2, \dots, A_{n-1})$$

27 March 2007

Statistics and Probability

Lec 2

Independence

- We often use this word informally, but it has a *specific formal meaning*
- Outcomes, or events, are *independent* if the occurrence of one does not change the *probability* of the occurrence of the other
- For independent events *A* and *B*,

$$P(A \text{ AND } B) = P(A) * P(B)$$

27 March 2007

Statistics and Probability

Lec 2

Example

- Let's go back to our example of tossing 2 dice, with *A* = sum is 8, *B* = 1st is 3
- Are *A* and *B* *independent*?
- $P(A \text{ AND } B) = ??$
- What if *A* = sum is 7?

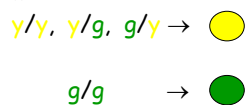
27 March 2007

Statistics and Probability

Lec 2

Mendel and peas

- Mendel's experiments with peas suggested to him that seed color (as well as other traits he examined) was caused by two different 'gene alleles' (he didn't use this terminology back then!)
- Each (non-sex) cell had two alleles, and these determined seed color:



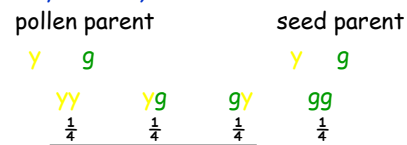
27 March 2007

Statistics and Probability

Lec 2

Peas, cont

- Here, yellow is dominant over green
- Sex cells each carry one allele
- Also postulated that the gene pair of a new seed was determined by the combination of its pollen and ovule, which are passed on *independently*



27 March 2007

Statistics and Probability

Lec 2

Conditional probability again

- Formally,



$$P(A | B) = P(A \text{ AND } B) / P(B)$$
- $P(A | B)$ is *not defined* if $P(B) = 0$
- Example:** Mendel and peas
 - Given that a seed is yellow, what is the probability it is a heterozygote?

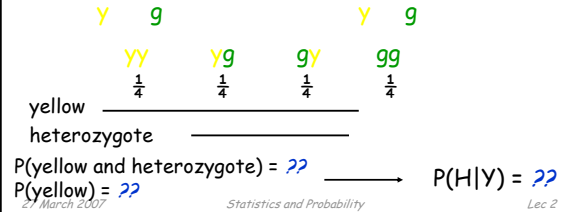
27 March 2007

Statistics and Probability

Lec 2

Mendel and peas again

- Heterozygotes $y/g, g/y$:
 - $y/y, y/g, g/y \rightarrow$ 
 - $g/g \rightarrow$ 
- pollen parent
- seed parent



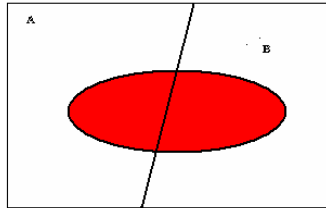
27 March 2007

Statistics and Probability

Lec 2

Partition

- A *partition* divides the sample space into disjoint subs
 - no gaps
 - no overlap.



27 March 2007

Statistics and Probability

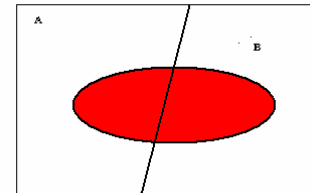
Lec 2

Law of total probability

- For a partition A, B of the sample space and an event R ,

$$P(R) = P(R \text{ AND } A) + P(R \text{ AND } B)$$

- Can have any number of events in the partition



27 March 2007

Statistics and Probability

Lec 2

Which box?

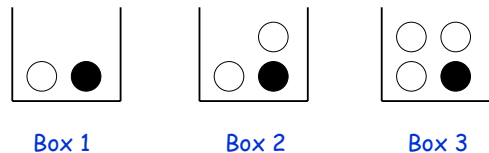
- Suppose there are 3 similar boxes: 1, 2, 3
- Box i contains i white balls and 1 black ball
- I choose a box at random and then pick 1 ball at random from that box (and show you the ball) - it is white
- Which box do you guess the ball came from??
- What is your chance of getting it right??

27 March 2007

Statistics and Probability

Lec 2

Which box? (cont)



27 March 2007

Statistics and Probability

Lec 2

Bayes' Rule

For a *partition* B_1, B_2, \dots, B_n of all possible outcomes,

$$\begin{aligned} P(B_i | A) &= P(A | B_i) * P(B_i) / P(A) \\ &= \frac{P(A | B_i) * P(B_i)}{P(A | B_1) * P(B_1) + \dots + P(A | B_n) * P(B_n)} \end{aligned}$$

27 March 2007

Statistics and Probability

Lec 2

Bayes' rule uses both expressions for $P(A \text{ AND } B)$

- Let's go back to the definition of conditional probability for a moment:
 $P(A | B) = P(A \text{ AND } B) / P(B)$
- We can also write
 $P(B | A) = P(A \text{ AND } B) / P(A)$
- This gives us *two ways* to express $P(A \text{ AND } B)$:
 $P(A \text{ AND } B) = P(A | B) P(B)$
 $P(A \text{ AND } B) = P(B | A) P(A)$

27 March 2007

Statistics and Probability

Lec 2

Example

- Suppose that 5% of men and 0.25% of women are color blind. A person is chosen at random. What is the probability that the person is *color blind* assuming...
 - there are an *equal number* of males and females
 - there are *twice as many males as females* in the population

27 March 2007

Statistics and Probability

Lec 2

Example (cont)

- Suppose that 5% of men and 0.25% of women are color blind. A *color blind* person is chosen at random. What is the probability that the person is *male* assuming...
 - there are an *equal number* of males and females
 - there are *twice as many males as females* in the population

27 March 2007

Statistics and Probability

Lec 2

Disease screening

- Suppose a blood test for a particular disease results in either a + or - reading
- Assume that 95% of people with the disease produce +, and 2% of people without the disease also produce +
- Also assume that 1% of the population has the disease
- What is the chance that a person chosen at random from the population has the disease, given that the test produces +?

27 March 2007

Statistics and Probability

Lec 2

Huntington's disease

- Rare, autosomal dominant disorder which starts to appear in middle age (40s or 50s)
- Incidence in the *general population* is about 1/10,000, which means that the probability of carrying the defective gene is about .01%
- But for individuals with an affected parent, the probability of carrying the defective gene is 50%
- There is a test for HD, assume that the false positive rate is 2%, and the false negative rate is 1%

27 March 2007

Statistics and Probability

Lec 2

Huntington's disease (cont)

- Suppose a person with an affected parent takes the test
 - What is the chance the person is at risk if the test is negative?
 - What is the chance the person is at risk if the test is positive?

27 March 2007

Statistics and Probability

Lec 2

Huntington's disease (cont)

- Now suppose that a person without an affected parent is considering taking the test
 - What is the chance the person is at risk if the test is negative?
 - What is the chance the person is at risk if the test is positive?

27 March 2007

Statistics and Probability

Lec 2