2004-09-30

STATS 3N03/3J04 8-1

"JARGON" AND "WHERE ARE WE?"
PAGES UPDATED AFTER EACH CLASS.

ASSIGNMENTS: SLOTS BY HH-105
MON 18:00

TEST #1 : MON 7-9PM, TUE 7-9PM, WED 8:30 AM
3N03 3J04 MAKEUP

MORE ON DATA FRAMES:

posf ← as.factor(baked$position)

- CREATES VECTOR IN LOCAL WORK SPACE, NOT IN DATA FRAME

baked$posf ← as.factor(position)

- USES OBJECT position FROM LOCAL WORKSPACE IF IT EXISTS, FAILS OTHERWISE.

```
> baked
   density position temp posf tempf
1      570        1  800    1   800
2      565        1  800    1   800
3      583        1  800    1   800
4     1063        1  825    1   825
5     1080        1  825    1   825
6     1043        1  825    1   825
7      565        1  850    1   850
8      510        1  850    1   850
9      590        1  850    1   850
10     528        2  800    2   800
11     547        2  800    2   800
12     521        2  800    2   800
13     988        2  825    2   825
14    1026        2  825    2   825
15    1004        2  825    2   825
16     526        2  850    2   850
17     538        2  850    2   850
18     532        2  850    2   850
> sapply(baked, is.factor)
 density position     temp     posf    tempf
   FALSE    FALSE    FALSE     TRUE     TRUE
> names(baked)
[1] "density"  "position" "temp"     "posf"     "tempf"
> names(baked)[2] <- "pos"
> names(baked)
[1] "density" "pos"     "temp"    "posf"    "tempf"
> rownames(baked)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10" "11" "12" "13" "14"
"15"
[16] "16" "17" "18"
```

```
> dim(baked)
[1] 18  5
> nrow(baked)
[1] 18
> ncol(baked)
[1] 5
> is.data.frame(baked)
[1] TRUE
> is.list(baked)
[1] TRUE
> is.matrix(baked)
[1] FALSE
> as.matrix(baked)
   density pos temp  posf tempf
1   " 570"  "1" "800" "1"  "800"
2   " 565"  "1" "800" "1"  "800"
3   " 583"  "1" "800" "1"  "800"
4   "1063"  "1" "825" "1"  "825"
5   "1080"  "1" "825" "1"  "825"
6   "1043"  "1" "825" "1"  "825"
7   " 565"  "1" "850" "1"  "850"
8   " 510"  "1" "850" "1"  "850"
9   " 590"  "1" "850" "1"  "850"
10  " 528"  "2" "800" "2"  "800"
11  " 547"  "2" "800" "2"  "800"
12  " 521"  "2" "800" "2"  "800"
13  " 988"  "2" "825" "2"  "825"
14 "1026"  "2" "825" "2"  "825"
15 "1004"  "2" "825" "2"  "825"
16  " 526"  "2" "850" "2"  "850"
17  " 538"  "2" "850" "2"  "850"
18  " 532"  "2" "850" "2"  "850"
```

```
> is.matrix(baked[, 1:3])
[1] FALSE
> as.matrix(baked[,1:3])
   density pos temp
1      570   1  800
2      565   1  800
3      583   1  800
4     1063   1  825
5     1080   1  825
6     1043   1  825
7      565   1  850
8      510   1  850
9      590   1  850
10     528   2  800
11     547   2  800
12     521   2  800
13     988   2  825
14    1026   2  825
15    1004   2  825
16     526   2  850
17     538   2  850
18     532   2  850
>
```

8-5

SOME DEFINITIONS...

Updated 2003-12-06

An experiment that can result in different outcomes, even though it is repeated in the same manner every time, is called a **random experiment**.

The set of all possible outcomes of a random experiment is called the **sample space** of the experiment.

An **event** is a subspace of the sample space of a random experiment.

A sample space is **discrete** if it consists of a finite (or countably infinite) set of outcomes.

Probability is a measure of certainty on a scale of 0 to 1. The probability of an impossible event is 0, the probability of an inevitable event is 1. If A and B are events, then P(A+B) = P(A) + P(B) - P(A.B), where A+B denotes set union and A.B denotes set intersection. Any one of the following three definitions of probability can be used to assign a probability to an event that is neither impossible nor inevitable.

> The **relative frequency definition of probability** applies when the sample space consists of elementary outcomes which, through physical symmetry, are recognized as being equally likely. The probability of an event E is the number elementary outcomes in E divided by the number of elementary outcomes in the sample space.
>
> The **limiting frequency definition of probability** applies when you can envisage a sequence of independent trials. Consider the number of trials that result in an event E, divided by the total number of trials. The probability of E is the hypothetical limit to which any such series of trials will tend.
>
> The **subjective definition of probability** defines your personal probability of an event E as the maximum amount of money you are willing to bet in order to win $1 if E occurs.

A **random variable** is a function that assigns a real number to each outcome in the sample space of a random experiment.

A **discrete random variable** is a random variable with a finite (or countably infinite) range.

A **continuous random variable** is a random variable with an interval (either finite or infinite) of real numbers for its range.

The **probability density function** for a random variable X is a non-negative function f(x) which gives the relative probability of each point x in the sample space of X. It integrates to 1 over the whole sample space. The integral over any subset of the sample space gives the probability that X will fall in that subset.

A **parameter** is a scalar or vector that indexes a family of probability distributions.

The **expected value** or **mean** or **average** of a random variable is computed as a sum (or integral) over all possible values of the random variable, each weighted by the probability of getting that value. It can be interpreted as the centre of mass of the probability distribution.

PROBABILITY

- WORK THROUGH ALL OF CHAPT. 2 IN DETAIL, DO SOME EXERCISES FROM EACH SECTION, MAKE SURE YOU UNDERSTAND IT.

- WANT A NUMBER BETWEEN 0 AND 1 TO DESCRIBE AN EVENT THAT IS NEITHER IMPOSSIBLE NOR INEVITABLE.

RELATIVE FREQUENCY DEFINITION:

- PHYSICAL SYMMETRY WITH "EQUALLY LIKELY" OUTCOMES

S : SET OF ALL POSSIBLE OUTCOMES

E ⊂ S : EVENT E

$$P(E) = \frac{\text{NO. OF OUTCOMES IN } E}{\text{NO. OF OUTCOMES IN } S}$$

EX "FAIR COIN" $S = \{H, T\}$, $E = \{H\}$

$$P(E) = \frac{1}{2}$$

EX "FAIR 6-SIDED DIE"

$S = \{1, 2, 3, 4, 5, 6\}$ "SAMPLE SPACE"

$E = \{2, 4\}$

$$P(E) = \frac{2}{6}$$

ADVANTAGE OF THIS DEFINITION

- UNAMBIGUOUS
- GIVES A PRECISE VALUE

DISADVANTAGE

- SELDOM APPLICABLE
- USE FOR GAMES OF CHANCE, MENDELIAN GENETICS, RANDOM SAMPLING FROM A GIVEN POPULATION

LIMITING FREQUENCY DEFINITION

IMAGINE A SEQUENCE OF n INDEPENDENT TRIALS OF THE CHANCE SET-UP

AFTER n TRIALS, DEFINE

$$P_n(E) = \frac{\text{NO. OF TRIALS RESULTING IN } E}{n}$$

$$P(E) = \lim_{n \to \infty} P_n(E)$$

EX ROLL A FAIR 6-SIDED DIE n TIMES

$$P(\{2,4\}) = \lim_{n \to \infty} \frac{(\text{NO. OF TIMES A 2 OR A 4 COMES UP})}{n}$$

ADVANTAGE
- DON'T NEED "EQUALLY LIKELY" ELEMENTARY OUTCOMES
- CAN ESTIMATE WITH, SAY, $n = 100$ TRIALS

DISADVANTAGE
- CAN'T GET EXACT VALUE

- CAN'T USE IF YOU CAN'T REPEAT THE EXPERIMENT

SUBJECTIVE DEFINITION

P(E) = THE MAXIMUM AMOUNT YOU ARE WILLING TO BET, TO WIN $1 IF E HAPPENS

ADVANTAGE

- ALWAYS APPLICABLE

DISADVANTAGE

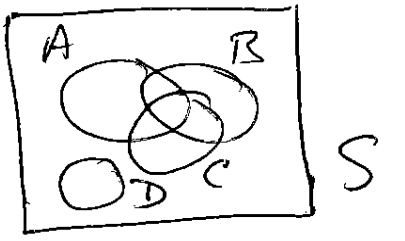
- DIFFERENT PEOPLE WILL HAVE DIFFERENT PROBABILITIES FOR THE SAME EVENT

- MAY ACHIEVE CONCENSUS IF INFORMATION IS POOLED AND DIFFERENT PEOPLE DISCUSS IT, OR MAYBE NOT.

AXIOMS OF PROBABILITY

$P(\phi) = 0$

$P(S) = 1$



$0 \leq P(A) \leq 1$ $\quad P(A \cup D) = P(A) + P(D)$

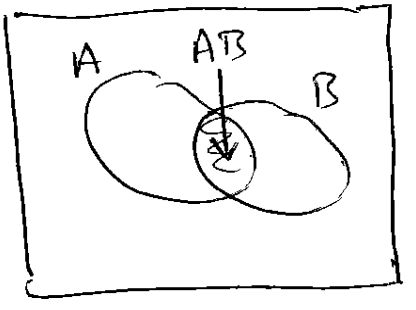
$P(A \cup B) = P(A) + P(B) - P(AB)$

"UNION"
"AND/OR"

$A \cdot B$
$A \cap B$
"INTERSECTION"
"AND"

DEFINITION OF CONDITIONAL PROBABILITY

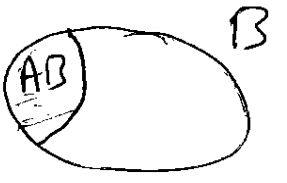
$$P(A|B) = \frac{P(AB)}{P(B)}$$



GIVE THAT

→

B HAPPENED



↑

B IS NOW THE WHOLE SAMPLE SPACE