ANIMAL GENETICS & BREEDING

UNIT – I

BIO-STATISTICS AND COMPUTER APPLICATION

Theory

PROBABILITY

Dr Anil Meel
Department of Animal Genetics & Breeding
MJF Veterinary college

Sample space and Events

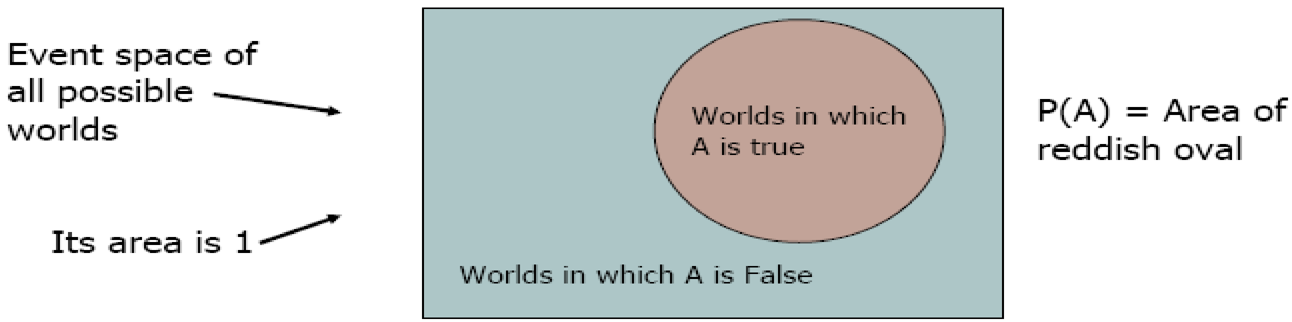
- Ω : Sample Space, result of an experiment
  - If you toss a coin twice $\Omega = \{HH, HT, TH, TT\}$
- Event: a subset of Ω
  - First toss is head = {HH,HT}
- S: event space, a set of events:
  - Closed under finite union and complements
    - Entails other binary operation: union, diff, etc.
  - Contains the empty event and Ω

Probability Measure

- Defined over (Ω, S) s.t.
  - $P(\alpha) >= 0$ for all α in S
  - $P(\Omega) = 1$
  - If α, β are disjoint, then
    - $P(\alpha \cup \beta) = p(\alpha) + p(\beta)$
- We can deduce other axioms from the above ones
  - Ex: $P(\alpha \cup \beta)$ for non-disjoint event

  $P(\alpha \cup \beta) = p(\alpha) + p(\beta) - p(\alpha \cap \beta)$

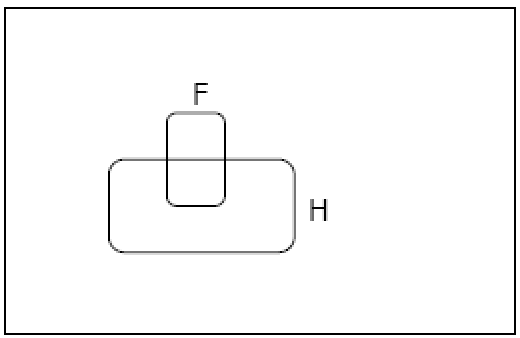
Visualization



- We can go on and define conditional probability, using the above visualization

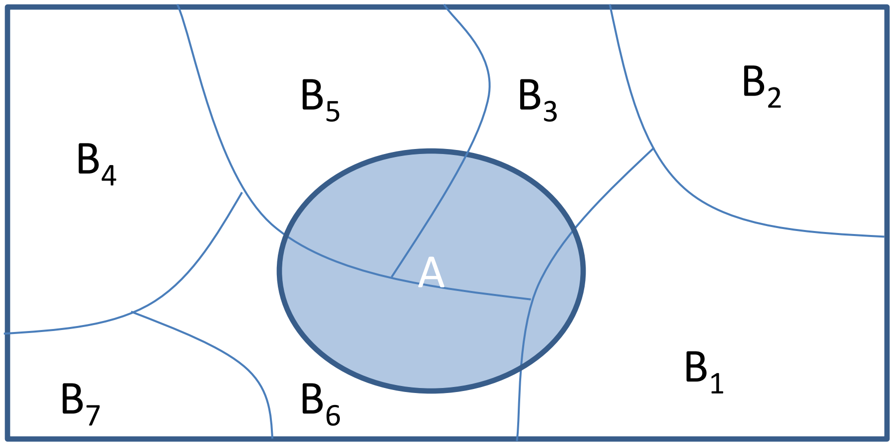
Conditional Probability

P(F|H) = Fraction of worlds in which H is true that also have F true



$$p(f \mid h) = \frac{p(F \cap H)}{p(H)}$$

Rule of total probability

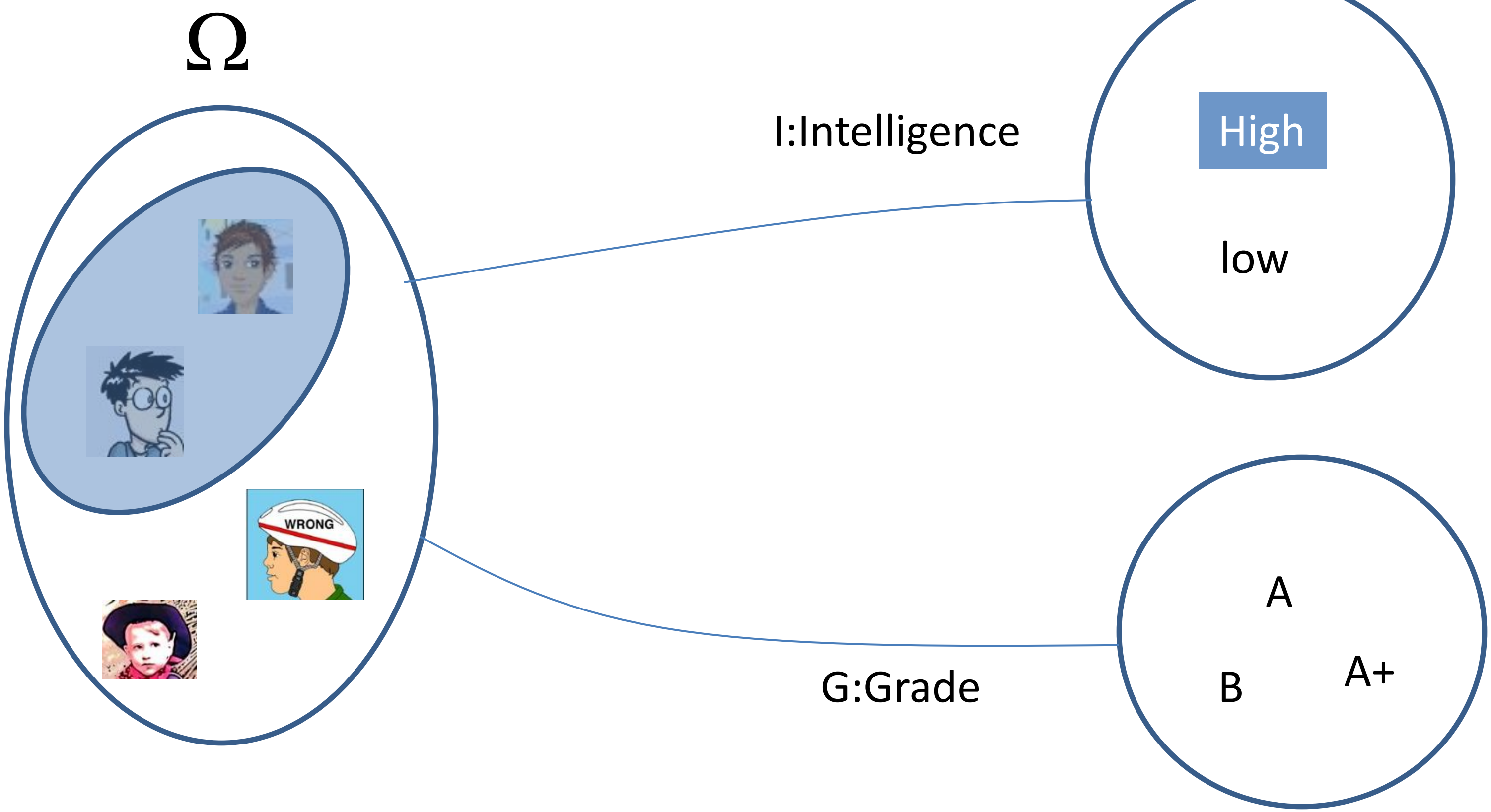


$$p(A) = \sum P(B_i)P(A \mid B_i)$$

From Events to Random Variable

- Almost all the semester we will be dealing with RV
- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - $\Omega =$ all possible students
  - What are events
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - Intelligence_High = … with high intelligence
  - Very cumbersome
  - We need “functions” that maps from Ω to an attribute space.
  - P(G = A) = P({student ϵ Ω : G(student) = A})

Random Variables



P(I = high) = P({all students whose intelligence is high})

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of tails X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2, …, 100

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
  - $\Sigma_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$

Common Distributions

- Uniform X $U[1, \ldots, N]$
  - X takes values 1, 2, … N
  - $P(X = i) = 1/N$
  - E.g. picking balls of different colors from a box
- Binomial X $Bin(n, p)$
  - X takes values 0, 1, …, n
  - $p(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$
  - E.g. coin flips

Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable x.

Probability of Continuous RV

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

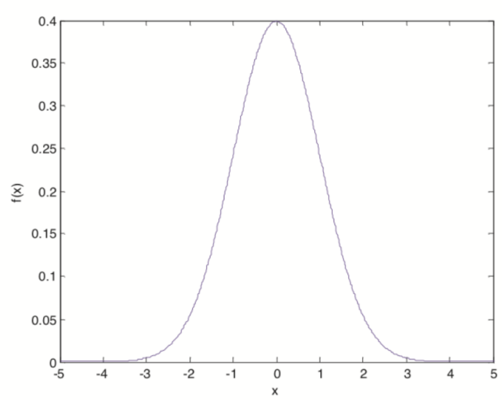
$$P(0 \leq X \leq 1) = \int_{0}^{1} f(x)dx$$

Cumulative Distribution Function

- $F_X(v) = P(X \leq v)$
- Discrete RVs
  - $F_X(v) = \Sigma_{vi}\, P(X = v_i)$
- Continuous RVs
  - $$F_X(v) = \int_{-\infty}^{v} f(x)dx$$
  - $$\frac{d}{dx} F_x(x) = f(x)$$

Common Distributions

- Normal X $\quad N(\mu, \sigma^2)$
  - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
  - E.g. the height of the entire population



Joint Probability Distribution

- Random variables encodes attributes
- Not all possible combination of attributes are equally likely
  - Joint probability distributions quantify this
- P(X= x, Y= y) = P(x, y)
  - Generalizes to N-RVs
  - $\sum_x \sum_y P(X = x, Y = y) = 1$
  - $\int_x \int_y f_{X,Y}(x, y) dxdy = 1$

Chain Rule

- Always true
  - $P(x, y, z) = p(x)\, p(y|x)\, p(z|x, y)$

    $= p(z)\, p(y|z)\, p(x|y, z)$

    $= \ldots$

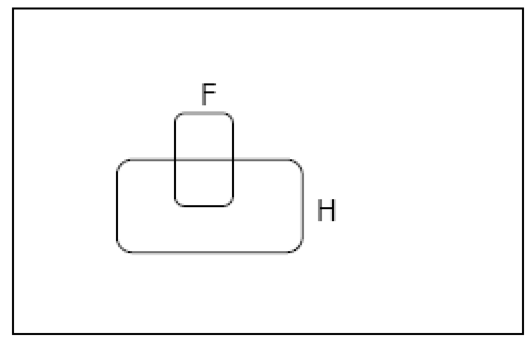
Conditional Probability

events

$$P\left(X = x \middle| Y = y\right) = \frac{P\left(X = x \cap Y = y\right)}{P\left(Y = y\right)}$$

But we will always write it this way:

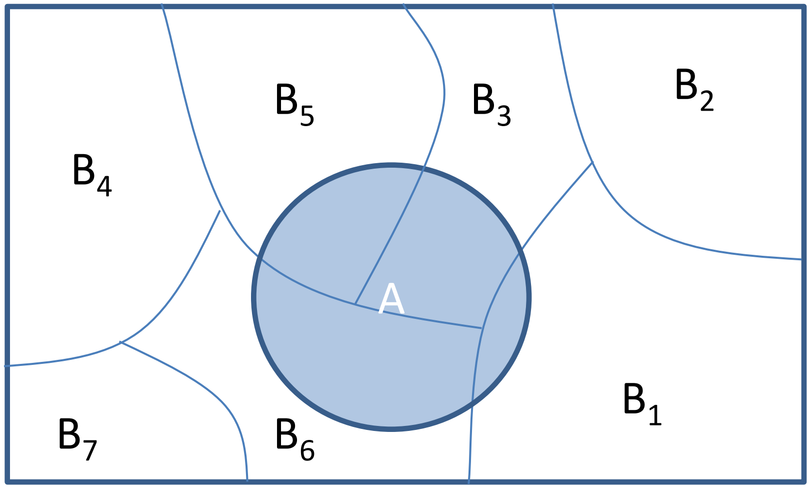
$$P(x \mid y) = \frac{p(x, y)}{p(y)}$$



Marginalization

- We know p(X, Y), what is P(X=x)?
- We can use the low of total probability, why?

$$p(x) = \sum_y P(x, y)$$
$$= \sum_y P(y) P(x \mid y)$$



Marginalization Cont.

- Another example

$$p(x) = \sum_{y,z} P(x, y, z)$$

$$= \sum_{z,y} P(y, z) P(x \mid y, z)$$

Bayes Rule

- We know that P(rain) = 0.5
  - If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(rain \mid wet) = \frac{P(rain)P(wet \mid rain)}{P(wet)}$$

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)}$$

Bayes Rule cont.

- You can condition on more variables

$$P(x \mid y, z) = \frac{P(x \mid z) P(y \mid x, z)}{P(y \mid z)}$$

Independence

- X is independent of Y means that knowing Y does not change our belief about X.
  - $P(X|Y=y) = P(X)$
  - $P(X=x, Y=y) = P(X=x)\, P(Y=y)$
  - The above should hold for all x, y
  - It is symmetric and written as $X \perp Y$

Independence

- $X_1, \ldots, X_n$ are independent if and only if

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} P\left(X_i \in A_i\right)$$

- If $X_1, \ldots, X_n$ are independent and identically distributed we say they are *iid* (or that they are a random sample) and we write

$$X_1, \ldots, X_n \sim P$$

CI: Conditional Independence

- RV are rarely independent but we can still leverage local structural properties like Conditional Independence.
- $X \perp Y \mid Z$ if once Z is observed, knowing the value of Y does not change our belief about X
  - P(rain $\perp$ sprinkler's on | cloudy)
  - P(rain $\not\perp$ sprinkler's on | wet grass)

Conditional Independence

- $P(X=x \mid Z=z, Y=y) = P(X=x \mid Z=z)$
- $P(Y=y \mid Z=z, X=x) = P(Y=y \mid Z=z)$
- $P(X=x, Y=y \mid Z=z) = P(X=x \mid Z=z)\, P(Y=y \mid Z=z)$

We call these factors : very useful concept !!

The Big Picture