Probability Theory – A "***very***" brief introduction

Based on Lectures by Andrew Moore and Xiaoli Fern

Notations

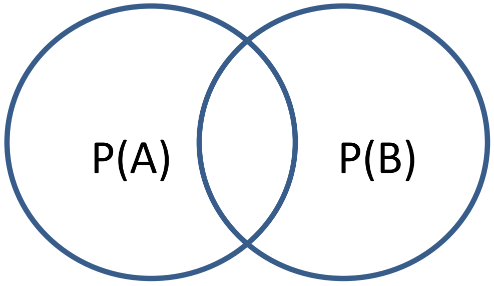
- Random Variable – An element whose value is not known or unobserved
- Example A = "It will rain today"
- Domain
  - Boolean (True/False)
  - Discrete (Grade of a student)
  - Continuous (The amount of rainfall)

Simple axioms

1. $0 \leq P(A) \leq 1$
2. $P(true) = 1$, $P(false) = 0$
3. $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$

If events A and B are independent

$$P(A \wedge B) = 0$$

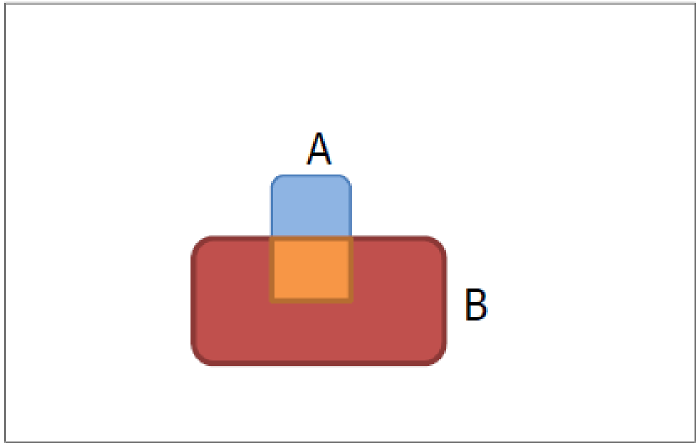


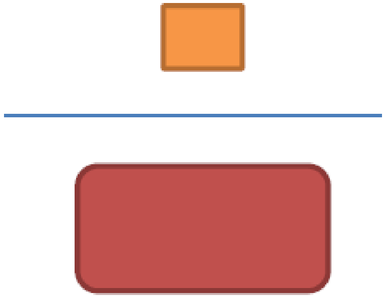
Joint Distribution

- Probability that a set of random variables “jointly” take a specific value combination
- Notation - P(A B) or P(A, B)
- Indicates that both A and B are true (in a boolean case)
- Example: P(high glucose, high BP)
- If A and B are independent then P(A,B) = P(A)P(B)

Conditional Probability

- P(A|B) = Fraction of worlds in which B is true that also have A true

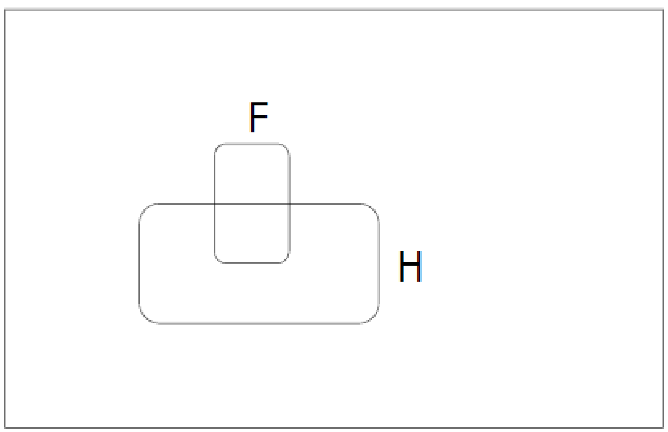




- If A and B are independent, P(A|B)=P(A)

Conditional Probability

- Some times, knowing one or more random variables can improve upon our prior belief of another random variable



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

"Headaches are rare (1/10), but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

Chain Rule

$$P(A,B,C,D) = P(D \mid C,B,A)P(C \mid B,A)P(B \mid A)P(A)$$

- More generally

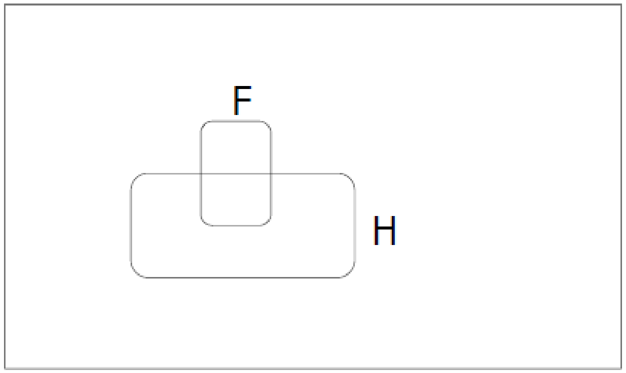
$$P(\cap_{k=1}^{n} A_k) = \prod_{1}^{n} P(A_k \mid \cap_{j=1}^{k-1} A_j)$$

- Key Application: Bayes Rule

$$P(A \wedge B) = P(A/B)\, P(B) = P(B/A)P(A)$$

$$P(A/B) = \frac{P(A \wedge B)}{P(B)}$$

Inference



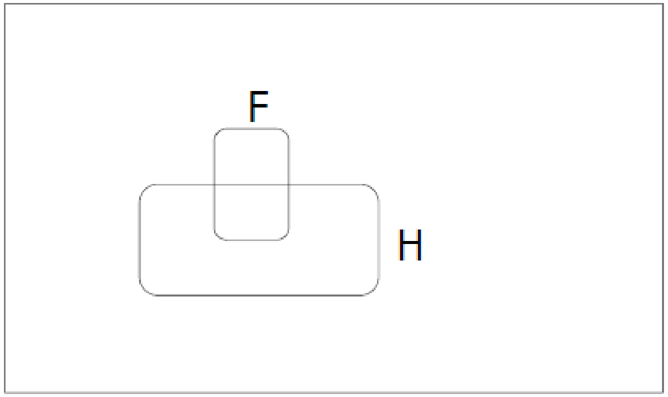
H = "Have a headache"
F = "Coming down with Flu"

$P(H) = 1/10$
$P(F) = 1/40$
$P(H|F) = 1/2$

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?

Inference



H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H|F) = 1/2

Prior: the degree of belief in an event in the absence of any other information

$$P(F|H) = \frac{P(F \wedge H)}{P(H)} = \frac{P(H|F)P(F)}{P(H)} = \frac{\frac{1}{40} * \frac{1}{2}}{1/10} = \frac{1}{8}$$

Posterior: the degree of belief in an event after obtaining some evidential information

More General Forms of Bayes Rule

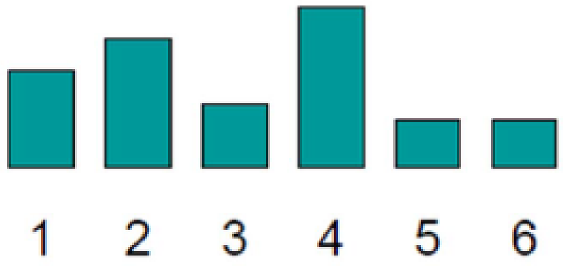
$$P(A|B\wedge X)=\frac{P(B|A\wedge X)P(A\wedge X)}{P(B\wedge X)}$$

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

$$P(A=v_i|B)=\frac{P(B|A=v_i)P(A=v_i)}{\sum_{k=1}^{n_A}P(B|A=v_k)P(A=v_k)}$$

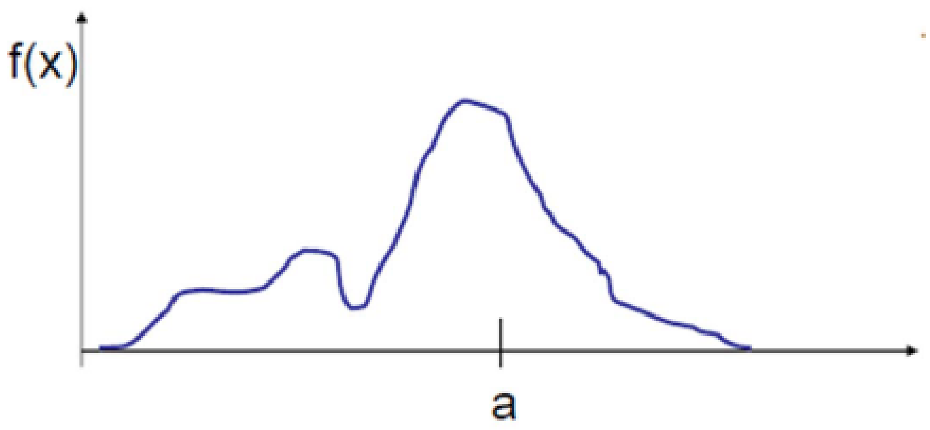
Probability distribution Functions

- Discrete distribution:



$$\sum_i P(X = x_i) = 1$$

- Continuous: Probability density function (PDF) f(x)

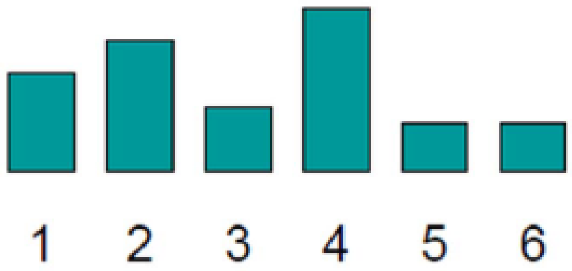


Multivariate

- Joint distribution of x and y is described by a **pdf** function f(x, y): $$P\left((x,y)\in A\right)=\int\int_A f(x,y)dxdy$$
- Marginal: $$f(x)=\int f(x,y)dy$$
- Conditional: $$f(x|y)=\frac{f(x,y)}{f(y)}$$
- Chain rule: $$f(x,y)=f(x|y)f(y)=f(y|x)f(x)$$
- Bayes rule: $$f(x|y)=\frac{f(y|x)f(x)}{f(y)}=\frac{f(y|x)f(x)}{\int f(y|x)f(x)dx}$$

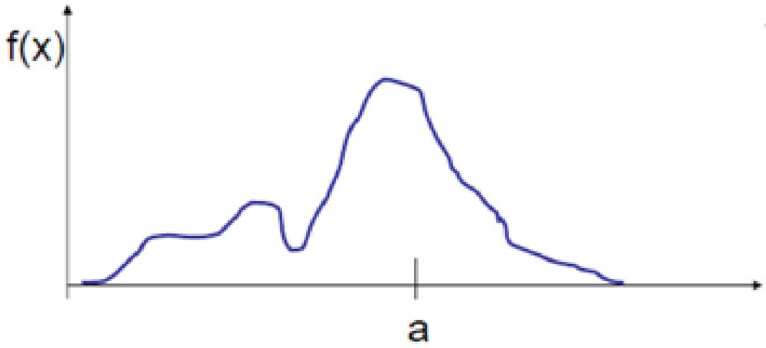
Expectations

- Expectation of a random variable of x is the weighted average of all possible values that x can take
- **Discrete :**



$$\bar{X} = E(X) = \sum_i x_i P(X = x_i)$$

- **Continuous:**



$$\bar{X} = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Variance

- Var(x) describes how far the values of x lie from the expected value of x (mean)

$$Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2$$

$$E[x^2] = \int x^2 f(x) dx$$

$$E[g(x)] = \int g(x) f(x) dx$$

Commonly Used Discrete Distributions

Bernoulli distribution: Ber(p)

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = p^x (1-p)^{1-x}$$

$$E(x) = p$$
$$Var(x) = p(1-p)$$

Binomial distribution: x ~ Binomial(n , p)

the probability to see x heads out of n flips

$$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$E(x) = np$$
$$Var(x) = np(1-p)$$

Gaussian (Normal)

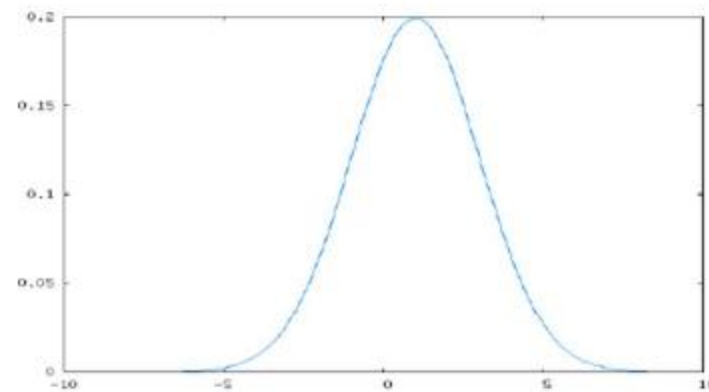
- If we look at the height of woman in the US, it will approximately look like Gaussian

$$x \sim N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

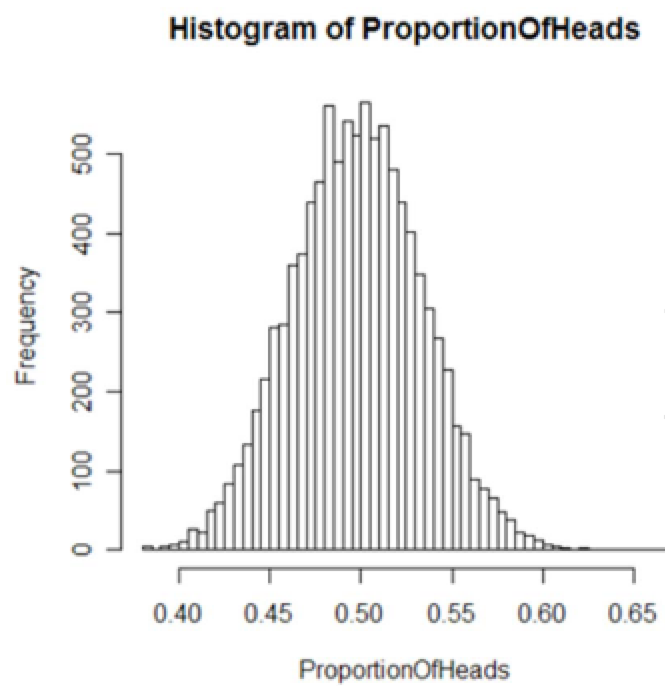
$$E[x] = \mu$$

$$Var(x) = \sigma^2$$



Central Limit Theorem

The sum of a large number of independent random variables is approximately Gaussian



average proportion of heads in a fair coin toss, over a large number of sequences of coin tosses