JOINS

→ inner join

→ left join

two dataframes A , B

A:

	a	b
index		
0		
1		
2		
3		
4		

B:

	c	d
index		
0		
2		

3

5

inner join: gives you rows with intersection

A inner join B

A.join(B, how='inner')

index values that are contained in both dataframes

index	a	b	c	d
0				
2				
3				

left join:

A.join(B, how='left')

	a	b	c	d
0			NA	NA
1				
2				
3				
4				

outer join

	a	b	c	d
0				
1				
2				
3				
4				
5	-	-	x	x
	1			

Probability

experiments

outcomes - "points" in the set of outcomes

events - sets of outcomes

toss a coin three times in a row

HHH, HHT, HTH, THH, HTT,

THT, TTH, TTT

simple events: {all three are heads}

= {HHH}

$E = \{\text{first toss is H}\}$

$= \{HHH, HHT, HTH, HTT\}$

$P(\cdot) \to$ maps events into $\mathbb{R}$.

(1) $P(E) \in [0,1]$

(2) $P(S) = 1$

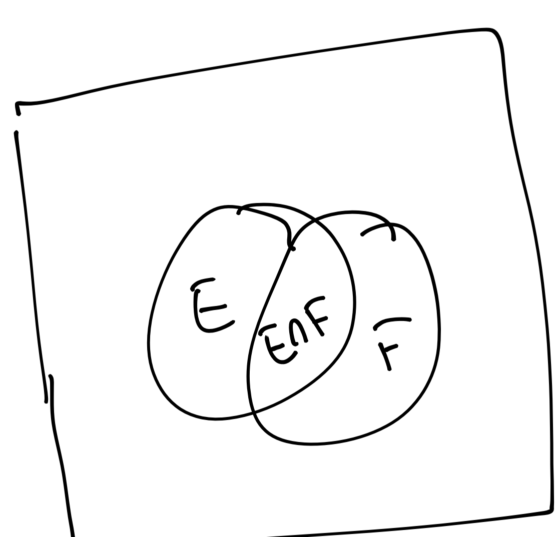
(3) $E_1, \ldots, E_n$ disjoints sets

$$E_i \cap E_j = \emptyset$$

$$P(\cup E_i) = \sum P(E_i)$$

E, F not necessarily disjoint

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$



$P(E|F)$ = probability that E occurs given that F has happened

$$= \frac{P(E \cap F)}{P(F)}$$

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

$E_1, \ldots, E_n$ disjoint $\qquad E, E^c$

$$P(A) = \sum_{i=1}^{n} P(A|E_i)P(E_i)$$

law of total probabilities

$$P(A) = P(A|E)P(E) + P(A|E^c)\underbrace{P(E^c)}_{1-P(E)}$$

Ex Test gives false positives 1%

false negatives 1%

P(really has infected | positive)

Random Variable

X with cdf F

random number where

$$P(X \leq x) = F(x)$$

two kinds of random variables:

1. discrete $\rightarrow$ finite / $\rightarrow$ countable
2. continuous

$\hookrightarrow$ take every value on at least one interval

discrete random variables have a pmf

$$p(x) = P(X = x)$$

$X \rightarrow \{0, 1, 2\}$

$p(0) = 0.5$
$p(1) = 0.2$
$p(2) = 0.3$

... cdf?

$\Rightarrow$ want to find the ...

$$F(x) = \begin{cases} 0, & x < 0 \\ 0.5, & 0 \le x < 1 \\ 0.7, & 1 \le x < 2 \\ 1.0, & x \ge 2 \end{cases}$$

continuous random variables has a pdf / density

$$f(x)$$

$$P(X \in [a,b]) = \int_a^b f(x)\,dx$$

$$\therefore \quad F(x) = \int_{-\infty}^{x} f(y)\,dy$$

$$f(x) = \begin{cases} 2x, & 0 \le x \le 1 \\ 0, & o/w \end{cases}$$

$x \in [0,1]$

$$F(x) = \int_{-\infty}^{x} f(y)\,dy$$

$$= \int_{-\infty}^{0} 0\,dy + 2\int_{0}^{x} y\,dy$$

$$= x^2 \qquad x \in [0,1]$$

$x > 1$

$$= \int_{-\infty}^{0} 0 \, dy + 2\int_{0}^{1} y \, dy$$

$$+ \int_{1}^{x} 0 \cdot dy = 1$$

Expectation: probability-weighted value

discrete random variables

$$E[X] = \sum_{p(x)>0} x \cdot p(x)$$

Ex
$p(0) = 0.5$

$p(1) = 0.2$

$p(2) = 0.3$

$$E[X] = 0 \cdot 0.5 + (1)(0.2)$$

$$+ (2)(0.3)$$

$$= 0.8$$

continuous random variable

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\,dx \qquad f(x) = \begin{cases} 2x, & 0<x<1 \\ 0, & o.w \end{cases}$$

$$E[X] = \int_{0}^{1} x \cdot 2x\,dx$$

$$= 2\int_{0}^{1} x^2\,dx = \frac{2}{3}$$

the expectation of a function of a rv $g(x)$

$$\Longrightarrow \underline{\sum g(x) \cdot p(x)} \quad \text{discrete}$$

$$\int_{-\infty}^{\infty} g(x) \cdot f(x)\,dx \quad \text{continuous}$$

$p(0) = .5$

$p(1) = .2$

$p(2) = .3$

$$g(x) = x^3 - 1$$

$$E[g(x)]$$

$$= (0^3 - 1)(.5)$$

$+ (1^3 - 1)(.2)$

$+ (2^3 - 1)(.3)$

$=$ a number! :)

$$\int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx$$ ((Continuous))

$$= 2\int_0^1 (x^3 - 1) \cdot x\, dx$$

$\Rightarrow$ Corollary $aX + b$

$$E[aX + b] = aE[X] + b$$

$$\mathrm{Var}(X) = E[(X - E(X))^2] = \sum (x - E[X])^2 \cdot p(x)$$

$\Rightarrow$ the average (probability weighted) squared distance from the expectation

$$= E[X^2] - \underbrace{E[X]^2}$$

$p(0) = .5$ $\quad E(X) = 0.8$

$p(1) = .2$ $\quad Var(X) = \underbrace{E[X^2]} - \underbrace{0.8^2}$

$p(2) = .3$

$$E[X^2] = 0^2 \cdot 0.5 + (1)^2(.2) + (2)^2(.3)$$

$$= 1.4$$

$$Var(X) = 1.4 - .64$$

$$Var(X) = E[(X - E[X])^2]$$

$$= E[X^2] - E[X]^2$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx$$

$$E[x^2] = 2\int_0^1 x^2 \cdot x\,dx$$

$$= 2\int_0^1 x^3\,dx$$

$$= \frac{1}{2}$$

X, Y both discrete

joint pmf

$$p(x,y) = P(X = x, Y = y)$$

$$E[g(x,y)] = \sum_y \sum_x g(x,y)\, p(x,y)$$

$X, Y \to \{0, 1\}$

- $p(0,0) = 0.5$
- $p(1,0) = 0.0$
- $p(0,1) = 0.4$
- $p(1,1) = 0.1$

$$E[x \cdot y] = 0 \cdot \cancel{0 \cdot p(0,0)} + (1) \cdot \cancel{0 \cdot p(1,0)} + 0 \cdot \cancel{1 \cdot p(0,1)} + (1)(1)\, p(1,1)$$

$$= 0.1$$

$f(x,y) \to$ joint density

$$P(X \in [a,b], Y \in [c,d])$$

$$= \int_c^d \int_a^b f(x,y)\,dx\,dy$$

$$E[g(x,y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\cdot f(x,y)\,dx\,dy$$

$$\text{Cov}(X,Y) = E[XY] - E[X]\cdot E[Y]$$

$p(0,0) = 0.5$

$p(1,0) = 0.0$

$p(0,1) = 0.4$

$p(1,1) = 0.1$

$E[X\cdot Y] = 0.1$

$E[X] = 0.1$

$E[Y] = 0.5$

$$\text{Cov}(X,Y) = 0.1 - (0.1)(0.5)$$

$$= .05$$

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{SD(X)}\sqrt{SD(Y)}}$$

$$SD(X) = \sqrt{Var(X)}$$

$\mathrm{Bern}(p)$ $\quad p \in [0,1]$

$f(0) = 1-p$

$f(1) = p$

$X \sim \mathrm{Bin}(n,p)$: sum of n independent $\mathrm{Bern}(p)$

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}$$

$\Longrightarrow$ X, Y are independent if for all sets

$$P(X \in A, Y \in B) = P(X \in A)\, P(Y \in B)$$

$$p(x,y) = p_X(x) \cdot p_Y(y)$$

$$f(x,y) = f_X(x) \cdot f_Y(y)$$

$$f(x,y) = cxy$$

$$f(x,y) = x+y$$

Normal random variable

$$E[X] = \mu$$

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

$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Central Limit Theorem

$X_1, \ldots, X_n$ independent identically distributed

$$E[X_i] = \mu$$

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

n large

$$\bar{X}_n \underset{\text{approx}}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$