**week 1 probs models and axioms**

sample space

- list(set) of possible outcomes,$\Omega$
- list must be:
  - mutually exclusive
  - collectively exhaustive
  - at the right granularity

prob axioms

- event:a subset of the smaple space-prob is assigned to event
- axioms:
  - nonnegative:$P(A) \geq 0$
  - normalization:$P(\Omega) = 1$
  - (finte) additivity:if $AB = \emptyset$, then$P(A \cup B) = P(A + B)$

some consequences of the axioms
if $A \subset B$,then $P(B) \geq P(A)$
$P(A \cup B) = P(A) + P(B) - P(AB)$
$P(AB) \leq P(A) + P(B)$
$P(A \cup B \cup C) = P(A) + P(A^c B) + P(A^c B^c C)$
discrete uniform law

- assume $\Omega$ consists of n equally likely elements
- assume A consist of k elements then $P(A) = \frac{k}{n}$

uniform prob law:prob=area
countable additivity axiom if $A_i$ is infinite sequence of disjoint events,then
$P(\cup_i A_i) = \sum_i P(A_i)$
de morgan's law
$(\cup_n S_n)^c = \cap_n S_n^c, (\cap_n S_n)^c = \cup_n S_n^c$
the geometric series $\sum_{i=0}^{\infty} \alpha_i = \frac{1}{1-\alpha}, |\alpha| \leq 1$
order of sum in series with multiple indices
$\sum_{i \geq 1, j \geq 1} a_{ij} = \sum_{i=1}^{\infty}(\sum_{j=1}^{\infty} a_{ij}) = \sum_{j=1}^{\infty}(\sum_{i=1}^{\infty} a_{ij})$

**week 2 conditioning and independence**

## conditioning and bayes' rule

conditional prob: $P(A|B)$ =prob of A, given that B occured
$P(A|B) = \frac{P(AB)}{P(B)}$ defined only when $P(B) \geq 0$
the multiplication rule
$P(AB) = P(A)P(B|A)$
$P(\cap_i A_i) = P(A_1) \prod_{i=2}^{n} P(A_i| \cap_{i-1} A_i)$
total prob theorem $P(B) = \sum_i P(A_i)P(B|A_i)$
bayes' rule $P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_j P(A_j)P(B|A_j)}$

## independent

independence of two events $P(AB) = P(A)P(B)$
conditional independence
conditonal independence, given C, is defined as independence under the prob law $P(.|C)$
$P(AB|C) = P(A|C)P(B|C)$
reliability

- chuan $p(chuan) = \prod_i p_i$
- bing $p(bing) = 1 - \prod_i (1 - p_i)$

**week3 counting**

discrete uniform law

- assume $\Omega$ consist of $n$ equally likely elements
- assume A consists of $k$ elements

then:$P(A) = \frac{\#A}{\#\Omega} = \frac{k}{n}$
combinations
def:$\binom{n}{k}$ numbers of k-elements subsets of a given n-elements sets $= \frac{n!}{k!(n-k)!}$
two ways of constructing an ordered sequence of k distinct items:

- choose the k items one at a time
- choose k items,then order them

useful formula
$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \binom{n}{n} = 1, \binom{n}{0} = 1, 0! = 1, \sum_{k=0}^{k} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + ... + \binom{n}{n} = \#$ all subsets $= 2^n$
binomial coefficient $\binom{n}{k} ->$ binomail probs

- $n \geq 1$ independent coin tosses;$P(H) = p$
- $P(HTTHHH) = p(1-p)(1-p)ppp = p^4(1-p)^2$
- $P(\text{particular sequence}) = p^{\#\text{ heads}}(1-p)^{\#\text{ tails}}$
- $P(particulark - headsequence) = p^k(1-p)^{n-k}$
- $P(heads) = \binom{n}{k} p^k (1-p)^{n-k} = p^k(1-p)^{n-k}.(\#$ k-head sequences)

partitions

- $n \geq 1$ distinct items,$r \geq 1$ persons given $n_i$ items to persion $i$
  - here $n_1, ..., n_r$ are given nonnegative integers
  - with $n_1 + ... + n_r = n$
- ordering n items:$n!$
  - deal $n_i$ to each person i, and then order

$cn_1!n_2!...n_r! = n!$ solve this formula we get number of partitions $\frac{n!}{n_1!n_2!...n_r!}$(multinomial coefficient)
the multinomial probs

- balls of different colors:$i = 1, ..., r$
- prob of picking a ball of color $i$ is $p_i$
- draw n balls, independently
- given nonnegative numbers $n_i$, with $n_1 + n_r + ... + n_r = n$
- find P($n_1$ balls of color 1,$n_2$ colors of color 2,..., $n_r$ balls of color r)
- special case $r = 2$;colors: head and tails

$P(\text{particular sequence of type}(n_1, n_2, ..., n_r)) = p_1^{n_1} p_2^{n_2} ... p_r^{n_r}$
sequence of type $(n_1, n_2, ..., n_r) ->$ partition of {1,2,...,n} into subsets of sizes $n_1, n_2, ..., n_r$
$P(\text{get type}(n_1, n_2, ..., n_r)) = \frac{n!}{n_1!n_2!...n_r!} p_1^{n_1} p_2^{n_2} ... p_r^{n_r}$

**week 4 discrete random variables**

## prob mass functions and expectations

pmf of a discrete r.v X

- it is the prob law or prob distribution of X
- if we fix some x, then "$X = x$" is an event

$p_X(x) = P(X = x) = P(\{\omega \in \Omega s.t. X(\omega) = x\})$
properties:$p_X(x) \geq 0, \sum_x p_X(x) = 1$
discrete uniform random variable;parameters a,b

- parameters a,b,$a \leq b$
- experiment:pick one of $a, a+1, ..., b$ at random;all equally likely
- smaple space:$\{a, a+1, ..., b\}$ $b - a + 1$ possible values
- random varible X:$X(\omega) = \omega$
- model of:compete ingnorance
- special case:$a = b$

binomial random variable;parameters:positive integer $n, n \in [0, 1]$

- experiment:n independent tosses of a coin with P(heads)=p
- smaple space: set of sequence o f H and T, of length n
- random variable X: number of heads observed
- model of:number of successes in a given number of independent trails
- $p_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$, for $k = 0, 1, ..., n$

geometric random varivable;parameters $p : 0 < p \leq 1$

- experiment:infinitely many independent tosses of a coin,P(heads)=p
- sample sapce: set of infinite sequences of H and T
- random X: number of tosses unitl the first heads
- model of: waiting times;number of trails unitl a successes

- $p_X(X=k)=(1-p)^k p$

expectation/mean of a random variable

- motivation:play a game 1000 times, random gain at each play describe by:
- average gain
- defintion:$E(X)=\sum_x p_X(x)$
- interpretation: average in large number of independent repetitions of the experiment
- caution: if we have an infinite sum, it needs to be well defined, we assume $\sum_x |x|p_X(x) \leq \infty$
- bernoulli:E(X)=p
- uniform:E(x)=$\frac{n}{2}=\frac{a+b}{2}$
- polulation average:E(X)=$\frac{1}{n}\sum_i x_i$

elmentary properties of expectations

- if $X \geq 0$, then $E(X) \geq 0$
- if $a \leq X \leq b$, then $a \leq E(X) \leq b$
- if $c$ is a constant,$E(c)=c$

the expected balue rule, for calculating $E(g(X))$

- let X be a r.v. and let $Y=g(X)$
- averaging over $y: E(Y)=\sum_y yp_Y(y)$
- averaging obver $x: E(g(X))=\sum_x g(x)P_X(x)$
- caution:in general,$E(g(X)) \neq g(E(X))$

linearity of expectation: $E(aX+b)=aE(X)+b$

## variance, conditioning on an event,multiple r.v.'s

variance– a measure of the spread of a pmf

- random variable X, with mena $\mu=E(X)$
- distance from the mean:$X-\mu$
- average distance from the mean:$E(X-\mu)=\mu-\mu=0$
- def:variance:$\mathrm{var}(X)=E((X-\mu)^2)$
- calculation,using the expected value rule, $E(g(X)=\sum_x g(x)p_X(x))=\sum_x (x-\mu)^2 p_X(x)$
- standard deviation:$\sigma_X=\sqrt{\mathrm{var}(X)}$

properties of the variance

- notation:$\mu=E(X)$
- $\mathrm{var}(aX+b)=a^2\mathrm{var}(X)$
- a useful formula:$\mathrm{var}(X)=E(X^2)-(E(X))^2$

variance of the bernoulli:$p(1-p)$

variance of the uniform:$\frac{1}{12}n(n+2)=\frac{1}{12}(b-a)(b-a+2)$

conditioning pmf and expectation, given an event

conditioning on an event A => use condional probs

$p_X(x)=P(X=x) \to p_{X|A}(x)=P(X=x|A)$

$\sum_x p_X(x)=1 \to \sum_x p_{X|A}(x)=1$

$E(X)=\sum_x xp_X(x) \to E(X|A)=\sum_x p_{X|A}(x)$

$E(g(X))=\sum_x g(x)p_X(x) \to E(g(X)|A)=\sum_x g(x)p_{X|A}(x)$

total expectation theorem

$p_X(x)=P(A_1)p_{X|A_1}(x)+...+P(A_n)p_{X|A_n}(x)$

$E(x)=P(A_1)E(X|A_1)+...+P(A_n)E(X|A_n)$

conditioning a geometric random varivable

$X$: number of independent coin tosses until first head:P(head)=p

$p_X(X=k)=(1-p)^{k-1}p, k=1,2,3,....$

conditioned on $X \geq 1, X-1$ is geometric with parameters p

memeoryless: number of remaining coin tosses, conditioned on trails in the first tosses, is geometric,with parameters p

the mean of the geometric:$\mu=\frac{1}{p}$

multiple random variables and joint pmfs

joint pmf:$p_{X,Y}=P(X=x,Y=y)$

properties:

- $\sum_x \sum_y p_{X,Y}(x,y)=1$
- $p_X=\sum_y p_{X,Y}(x,y)$
- $p_Y=\sum_x p_{X,Y}(x,y)$

more than two random variables

$p_{X,Y,Z}=P(X=x,Y=y,Z=z)$

- $\sum_x \sum_y \sum_z p_{X,Y,Z}(x,y,z)=1$
- $p_X(x)=\sum_y \sum_z p_{X,Y,Z}(x,y,z)$
- $p_{X,Y}(x,y)=\sum_z p_{X,Y,Z}(x,y,z)$

functions of multiple random variables

- expected value rule:$E(g(X,Y))=\sum_x \sum_y g(x,y)p_{X,Y}(x,y)$
- linearity of expectations:$E(aX+b)=aE(X)+b, E(X+Y)=E(X)+E(Y)$

the mean of the binomial $\mu=np$

## conditioning on a random variable; independent of r.v.'s

conditional pmfs

$p_{X|Y}(x|y)=\frac{p_{X,Y}(x,y)}{p_Y(y)}$ defined for $y$ such that $p_Y(y) \geq 0$

conditional pmfs involving more than two random variables

- self-explanatory notation:$p_{X|Y,Z}(x|y,z)=\frac{p_{X,Y,Z}(x,y,z)}{p_{Y,Z}(y,z)}$
- $p_{X,Y|Z}(x,y|z)=P(X=x,Y=y|Z=z)$
- multiplication rule: $P(ABC)=P(A)P(B|A)P(C|AB) \to p_{X,Y,Z}(x,y,z)=p_X p_{Y|X}(y|x)p_{Z|X,Y}(z|x,y)$

conditional expectation

$E(X|A)=\sum_x xp_{X|X|A}(x|A)$

$E(g(X)|A)=\sum_x g(x)p_{X|A}(x|A)$

total prob and expectation theorem

$E(X)=\sum_y p_Y(y)E(X|Y=y)$

independence

X,Y,Z are independent if $p_{X,Y,Z}(x,y,z)=p_X(x)p_Y(y)p_Z(z)$ for all $x,y,z$

if X, Y is independent:$E(XY)=E(X)E(Y), \mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)$

$g(X),h(Y)$ are also independent:$E(g(X)h(Y))=E(g(X))E(h(Y))$

variance of the binomial:$\sigma^2=npq=np(1-p)$

the hat problem

- n people throw their hat in a box and then pick one at random
  - all permutations equally likely
  - equivalent to picking one hat at a time
- X: number of people who get their own hat
  - find E(X)=1
  - $X_i$=1, if selects own hat,0, otherwise
  - $X=X_1+...+X_n$
- $E(X_i)=E(X_1)=\frac{1}{n}$

the variance in the hat problem

- X: number of people who get their own hat
- find var(X)
- var(X)=$E(X^2)-(E(X))^2$
- $E(X_i^2)=E(X_1^2)=E(X_1)=1/n, X^2=\sum_i X_i^2+\sum_{i,j:i\neq j} X_iX_j, E(X^2)=n \times \frac{1}{n}+n(n-1)\frac{1}{n}\frac{1}{n-1}$
- for $i \neq j: E(X_iX_j)=E(X_1,X_2)=P(X_1X_2=1)=P(X_1=1,X_2=1)=P(X_1=1)P(X_2|X_1=1)=\frac{1}{n}\frac{1}{n-1}$

### week 5 contiuous random variables

## prob density functions

prob density functions-pdfs def: a random variable is continuous if it can be described by a pdf

$P(a \leq X \leq a+\delta) \simeq f_X(a).\delta$

$P(a \leq X \leq b)=\int_a^b f_X(x)dx$

$f_X(x) \geq 0$

$\int_{-\infty}^{\infty} f_X(x)dx=1$

expectation/mean of a continuous random variable

interpretation: average in large number of independent repetitions of the experiment

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

properties of expectation

- if $X \geq 0$,then $E(X) \geq 0$
- if $a \leq X \leq b$, then $a \leq E(X) \leq b$
- expected value rule:$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$
- linearity:$E(aX+b) = aE(X)+b$

variance and its properties

- def: var(X)=$E((X-\mu)^2)$
- caculation using the expected value rule:
- var(X)=$\int_{-\infty}^{\infty}(x-\mu)^2 dx$
- standard deviation:$\sigma_X = \sqrt{\text{var}(X)}$
- var$(aX+b)$=$a^2$var$(X)$
- useful fromula:var(X)=$E(X^2)-(E(X))^2$

uniform(a,b):

- $\mu = \frac{a+b}{2}$
- $\sigma^2 = \frac{(b-a)^2}{12}$

exponential random variable,parameter $\lambda > 0$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases}$$

- $E(X) = \frac{1}{\lambda}$
- $E(X^2) = \frac{2}{\lambda^2}$
- var(X) $= \frac{1}{\lambda^2}$

cumulative distribution function(cdf)

def:$F_X(x) = P(X \leq x)$

continuous random variable $F_X(x) = \int_{-\infty}^{x} f_X(t) dt$

$\frac{dF_X(x)}{dx}(x) = f_X(x)$

discrete random variables:$F_X(x) = P(X \leq x) = \sum_{k \leq x} p_X(k)$

general cdf properties

- non-decreasing, if $y \geq x, F_X(y) \leq F_X(x)$
- $F_X(x)$ tends to 1, as $x \to \infty$
- $F_X(x)$ tends to 0, as $x \to -\infty$

normal(gaussian) random variable

- important in the theory of prob - central limit theorem
- prevalent in applications
  - conveninent analytical properties
  - model fo noise consisting of many, small independent noise terms

standard normal random variables

- standard normal $N(0,1) : f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$
- $\int_{-\infty}^{\infty} e^{-x^2/2} = \sqrt{2\pi}$
- $\mu = 0$
- $\sigma = 1$

general normal random variable

- general normal $N(\mu, \sigma) : f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $E(X) = \mu$
- var(X)=$\sigma^2$

linear functions of a normal random variable

- let $Y = aX+b, X \sim N(\mu, \sigma^2), E(X) = a\mu + b, \text{var}(X) = a^2\sigma^2$
- fact $Y \sim N(a\mu+b, a^2\sigma^2)$

standardizing a random variable

- let X have mean $\mu$ and variance $\sigma^2 > 0$
- let $Y = \frac{X-\mu}{\sigma}$
- if also X is a normal, then $Y \sim N(0,1)$

## conditioning on an event; multiple r.v.'s

conditional pdfs,given an event

for $P(A) > 0$

- $f_X(x)\delta \simeq P(x \leq X \leq x+\delta)$
- $f_{X|A}(x)\delta \simeq P(x \leq X \leq x+\delta | A)$
- $P(X \in B) = \int_B f_X(x) dx$
- $P(X \in B | A) = \int_B f_{X|A}(x|A) dx$
- $\int f_{X|A}(x|A) dx = 1$

conditional pdf of X, given that $X \in A$

$$f_{X|X\in A}(x) = \begin{cases} 0, x \notin A, \\ \frac{f_X}{P(A)}, x \in A \end{cases}$$

conditional expectation of X, given an event

- $E(X|A) = \int x f_{X|A}(x) dx$
- $E(g(X)|A) = \int g(x) f_{X|A} dx$

joint continous r.v.'s and joint pdfs

- def: two random variable are jointly continous if they can be described by a joint pdf
- $f_{X,Y}(x,y) \geq 0$
- $P((X,Y) \in B) = \int\int_{(x,y)\in B} f_{X,Y}(x,y) dxdy$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y) = 1$

on joint pdfs

$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x,y) dxdy$

$P(a \leq X \leq a+\delta, c \leq Y \leq c+\delta) \simeq f_{X,Y}(a,c)\delta^2$

$f_{X,Y}(x,y)$ : prob per unit area

aera(B)=0,$\to P((X,Y) \in B) = 0$

from the joint to the marginals

$f_X(x) = \int f_{X,Y}(x,y) dy$

$f_Y(y) = \int f_{X,Y}(x,y) dx$

uniform joint pdf on a set S

$f_{X,Y}(x,y) = \{ \; \frac{1}{\text{area of} S}, (x,y) \in S 0, \text{otherwise}$

the joint cdf

$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}$

### conditioning on a random variable;independence;bayes rules

conditional pdfs, given another r.v. $f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{(f_Y(y))}, f_Y(y) > 0$

def:$P(X \in A | Y = y) = \int_A f_{X|Y}(x|y) dx$

comments on conditional pdfs

- $f_{X|Y}(x|y) \geq 0$
- think of value of Y as fixed at some y shape of $f_{X|Y}(.|y)$ : slice of the joint
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X|Y}(x|y) dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx}{f_Y(y)} = 1$
- multiplication rule:$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y) = f_X(x) f_{Y|X}(y|x)$

total prob and expectation theorems

- $f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y) dy$
- $E(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx$
- $E(X) = \int_{-\infty}^{\infty} f_Y(y) E(X|Y=y) dy$
- expected value rule: $E(g(X)|Y = y) = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx$
- independence:$f_{X,Y}(x,y) = f_X(x) f_Y(y), \text{for all} x, y$

### week 6 further topics on random variables

## derived distribution

a linear function of a discrete r.v.

$Y = aX + b : p_Y(y) = p_X(\frac{y-b}{a})$

a linear function of a continuous r.v.

$Y = aX + b : f_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$

a linear function of a continuous r.v. is normal

if $X \sim N(\mu, \sigma^2), \text{then} aX + b \sim N(a\mu + b, a^2\sigma^2)$

a general function g(X) of a continuous r.v.

two-step procedure:

- find the cdf of Y:$F_Y(y) = P(Y \leq y) = P(g(Y) \leq y)$
- differentiate:$f_Y(y) = \frac{dF_Y(y)}{dy}$

a general formula for the pdf of $Y = g(X)$ when g is monotonic

$f_Y(y) = f_X(h(y)) |\frac{dh(y)}{dy}|$

## sums of independent r.v.'s, covariance and correlation

the distribution of X+Y:the discrete case

$Z = X + Y$;X,Y independent, discrete $g(X, Y)$ known pmfs

$p_Z(z) = \sum_x p_X(x)p_Y(z - x)$

if the continuous case: $f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx$

the sum of finitely many independent normals is normal

covariance $cov(X, Y) = E((X - E(X))(Y - E(Y)))$

independent $\rightarrow cov(X, Y) = 0$

covariance properties

$cov(X, X) = var(X) = E(X^2) - (E(X))^2$

$cov(aX + b, Y) = acov(X, Y)$

$cov(X, Y + Z) = cov(X, Y) + cov(X, Z)$

$cov(X, Y) = E(XY) - E(X)E(Y)$

the variance of a sum of random variables

- $var(X_1 + X_2) = var(X_1) + var(X_2) + 2cov(X_1, X_2)$
- $var(\sum_i^n X_i) = \sum_i^n var(X_i) + \sum_{(i,j):i\neq j} cov(X_i, X_j)$

the correlation coefficient

- dimensionless version of covariance:$\rho(X, Y) = E(\frac{(X-E(X))}{\sigma_X}.\frac{Y-E(Y)}{\sigma_Y})$
- slope $-1 \leq \rho \leq 1$
- meansure of the degree of 'association' between X and Y
- independent $\rightarrow \rho = 0$, uncorrelated,converse is not true
- $\rho(X, X) = 1$
- $|\rho| = 1, <=> (X - E(X)) = c(Y - E(Y))$ (linearly related)
- $cov(aX + b, Y) = acov(X, Y) \rightarrow \rho(aX + b, Y) = sign(a)\rho(X, Y)$

## conditional expectation and variance revisited, sum of a random number of independent r.v.'s

conditional expectation as a random variable

- function h,$h(x) = x^2$
- random variable X, what is h(X)?
- h(X) is the r.v. that takes the value $x^2$,if X happens to take the value x
- $g(y) = E(X|Y = y) = \sum_x p_{X|Y}(x, y)$ (integral in continous case)
- $g(Y)$ is the r.v. that takes the value $E(X|Y = y)$, if Y happens to take the value y
- definition:$E(X|Y) = g(Y)$
- remarkes:
  - it is a function of Y
  - it is a random variable
  - has a distribution, mean , variance,etc

the mean of $E(X|Y)$:the law of iterated expectation $E(E(X|Y)) = E(X)$

forecast revisions

- suppose forecasts are made by calculating expected value, given any available information
- X: february sales
- forecast in the beginning of the year
- end of january: will get new information,value y of Y,revisited$E(X|Y = y)$
- law of iterated expectation E(revised forecast)=E(X)= orginal forecast

$var(X) = E((X - E(X))^2) \rightarrow var(X|Y) = E((X - E(X|Y = y))^2|Y = y)$

$var(X|Y)$ is r.v that takes the value var(X|Y=y),when Y=y

law of total variance $var(X) = E(var(X|Y)) + var(E(X|Y))$

derivate of the law of total variance

$var(X|Y = y) = E(X^2|Y = y) - (E(X|Y = y))^2$ for all y

$var(X|Y) = E(X^2|Y) - (E(X|Y))^2$

$E(var(X|Y)) = E(X^2) - E((E(X|Y))^2)$

$var(E(X|Y)) = E((E(X|Y))^2) - (E(E(X|Y)))^2$

section means and variance

var(X)= (average variable within sections)+(variable between sections)

sum of a random number of independent r.v.'s

- N: number of stores visited
- $X_i$: money spent in store i,$X_i$ independent,identically distributed,independent of N
- let $Y = X_1 + ... + X_N$
- $E(Y|N = n) = nE(X)$
- total expectation theorem:$E(Y) = \sum_n p_N(n)E(Y|N = n) = E(N)E(X)$
- law of iterated expectation:$E(Y) = E(E(Y|N)) = E(N)E(X)$

variance of sum of a random number of independentr.v.'s

$Y = X_1 + ... + X_N$

- $var(Y) = E(var(Y|N)) + var(E(Y|N))$
- $var(Y) = E(N)var(X) + (E(X))^2var(N)$
- $E(Y|N) = nE(X)$
- $var(Y|N) = Nvar(X)$
- $E(var(Y|N)) = E(N)var(X)$

### week 7 bayesian inference

## introduction to bayesian inference

the bayesian inference framework

- unknown $\Theta$
  - treated as a random variable
  - prior distribution $p_\Theta$ or $f_\Theta$
- observation X, observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- use appropriate version of the bayes rule to find $p_{\Theta|X}(.|X = x)$ or $f_{\Theta|X}(.|X = x)$

the output of bayesian inference

the complete answer is a posterior distribution:pmf $p_{\Theta|X}(.|x)$ or pdf $f_{\Theta|X}(.|x)$

point estimates in bayesian inference

estimate:$\hat{\theta} = g(x)$

estimator:$\hat{\Theta} = g(X)$

maximum a posterior prob(map):

$p_{\Theta|X}(\theta^*|x) = \max_\theta p_{\Theta|X}(\theta|x)$

$f_{\Theta|X}(\theta^*|x) = \max_\theta f_{\Theta|X}(\theta|x)$

least mean square(lms):conditional expectation $E[\Theta|X = x]$

discrete,$\Theta$,discrete X

- $p_{\Theta|X}(\theta|x) = \frac{p_\Theta(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$
- $p_X(x) = \sum_{\theta'} p_\Theta(\theta')p_{X|\Theta}(x|\theta')$

continuous $\Theta$,continuous X:

- $f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)f_{X|\Theta}(x|\theta)}{f_X(x)}$
- $f_X(x) = \int f_\Theta(\theta')f_{X|\Theta}(x|\theta')d\theta'$

inferrng the unknown bias of a coin and the beta distribution

- standard example
  - coin with bais $\Theta$;prior $f_\Theta(.)$
  - fix $n, K$=number of heads
- assume $f_\Theta(.)$ is uniform in $[0, 1]$
- $f_{\Theta|K}(\theta|k) = \frac{1.\binom{n}{k}\theta^k(1-\theta)^{n-k}}{p_k(k)} = \frac{1}{d(n,k)}\theta^k(1 - \theta)^{n-k}$,beta distribution,with parameters $(k + 1, n - k + 1)$,$\theta \in [0, 1]$
- if prior is beta,$f_\Theta(\theta) = \frac{1}{c}\theta^\alpha(1 - \theta)^\beta, \alpha, \beta \geq 0$
- $f_{\Theta|K}(\theta|k) = \frac{1}{c}\theta^\alpha(1 - \theta)^\beta\binom{n}{k}\theta^k(1 - \theta)^{n-k}/p_K(k) = d\theta^{\alpha+k}(1 - \theta)^{\beta+n-k}$
- $\hat{\theta} = k/n$
- $\hat{\Theta} = K/n$
- $\int_0^1 \theta^\alpha(1 - \theta)^\beta d\theta = \frac{\alpha!\beta!}{(\alpha+\beta+1)!}, \alpha, \beta \geq 0$
- $E(\Theta|K = k) = \int_0^1 \theta f_{\Theta|K}(\theta|k)d\theta = \frac{k+1}{n+2} \rightarrow k/n, ask, n \rightarrow large$

## linear model with normal noise

$X_i = \sum_{j=1}^{m} a_{ij}\Theta_j + W_i, W_i, \Theta_j$ : independent,normal

- very common and convenient model
- bayes' rule: normal posterior
- map and lms estimates coincide - simple formulas(linear in the observation)
- many nice properties
- trajectory estimation example

$f_X(x) = c.e^{-(\alpha x^2+\beta x+\gamma)}, \alpha > 0$,nomral with mean $\frac{-\beta}{2\alpha}$ and variance $\frac{1}{2\alpha}$

the case of multiple observation $\hat{\theta_{map}} = \hat{\theta_{lms}} = E(\Theta|X=x) = \frac{\sum_{i=0}^{n}\frac{x_i}{\sigma_i^2}}{\sum_{i=0}^{n}\frac{1}{\sigma_i^2}}$

- key conclusions:
  - posterior is normal
  - lms and map estimate conincide
  - these estimates are 'linear', of the form $\hat{\theta} = a_0 + a_1x_1 + ... + a_nx_n$
- interpretations:
  - estimate $\hat{\theta}$:weighted average of $x_0$ piror mean and $x_i$ observation
  - weights determined by variances

the mean squared error $E((\Theta - \hat{\Theta})^2|X=x) = E((\Theta - \hat{\Theta})^2) = \frac{1}{\sum_{i=0}^{n}\frac{1}{\sigma_i^2}}$

## least mean square estimation

lms estimation in the absence of observation

- minimize mean squared error,$E((\Theta - \hat{\theta})^2) : \hat{\theta} = E(\Theta)$
- optimal mean squared error:$E((\Theta - E(\Theta))^2) = var(\Theta)$

properties of the estimation error in lms estimation

- estimator $\hat{\Theta} = E(\Theta|X)$
- error $\tilde{\Theta} = \hat{\Theta} - \Theta$
- $E(\tilde{\Theta}|X=x) = 0$
- $cov(\tilde{\Theta}, \hat{\Theta}) = 0$
- $var(\Theta) = var(\hat{\Theta}) + var(\tilde{\Theta})$

### week 8 limit theorems and classical statistics

## inequality,convergence, and the weak law of large numbers

inequality

- bound $P(X \geq a)$ based on limited information about a distribution
- markov inequality based on mean
- chebyshev inequality based on the mean and variance

wlln:$X, X_1, ..., X_n, i.i.d$ $\frac{X_1+...+X_n}{n} \to E(X)$

- application to polling

precise defn. of convergence - convergence 'in prob'

the markov inequality

- use a bit of information about a distribution to learn sth about probs of 'extreme events'
- if $X \geq 0, E(X)$ is small, then X is unlikely to be very large

def: if $X \geq 0$ and $a > 0$, then $P(X \geq a) \leq \frac{E(X)}{a}$

the chebyshev inequality

- random variable X, with finite mean and variance
- if the variance is small, then X is unlikely to be too far from the mean

math forumula $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$

the weak law of large numbers(wlln)

- $X_1, ..., X_n$ i.i.d.; finite mean and variance
- sample mean $M_n = \frac{X_1+...+X_n}{n}$
- $E[M_n] = \mu$
- var$(M_n)=\frac{\sigma^2}{n}$
- $P(|M_n - \mu| \geq \epsilon) \leq var(M_n)/\epsilon^2 = \frac{\sigma^2}{n\epsilon^2}$

convergence in prob def: a sequence $Y_n$ converges in prob to a number $a$ if :for any $\epsilon > 0$, $\lim_{n\to\infty} P(|Y - a| \geq \epsilon) = 0$

### the central limit theorem

the central limit theorem

- $X_1, ..., X_n$ i.i.d., finite mean and variance
- $S_n = X_1 + ... + X_n$, variance:$n\sigma^2$
- $\frac{S_n}{\sqrt{n}} \to \sigma^2$
- let Z be a standard normal r.v. (zero mean,unit variance)
- clt:for every z:$\lim_{n\to\infty} P(Z_n \leq z) = P(Z \leq z)$
- $P(Z \leq z)$ is the standard normal cdf, $\Phi(z)$, available from the normal tables

usefulness of the clt

- universal and easy to apply; only means, variances matter
- fairly accurate computational shortcut
- justification of normal models
- $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$

## an introduction to classical statistics

estimating a mean

$X_1, ..., X_n$:i.i.d., mean and variance
$\hat{\Theta_n}$ =sample mean $= M_n = \frac{X_1+...+X_n}{n}$: estimator(a r.v.)

properties and terminology

- $E[\hat{\Theta_n}] = \theta$(unbiased)
- wlln:$\hat{\Theta_n} \to \theta$(consistency)
- mean squared error:$E((\hat{\Theta_n} - \theta)^2) = \frac{\sigma^2}{n}$
- for any estimator,using $E(Z^2) = var(Z) + (E(Z))^2$;$E[(\hat{\Theta_n} - \theta)^2] = var(\hat{\Theta}) + (\text{bias})^2$
- $\sqrt{var(\hat{\Theta})}$ is called the standard error

ci for the estimation of the mean

$P(\hat{\Theta_n} - \frac{1.96\sigma}{\sqrt{n}} \leq \theta \leq \hat{\Theta_n} + \frac{1.96\sigma}{\sqrt{n}}) \simeq 1 - \alpha = 0.95$

maximum likelihood estimation

- pcik $\theta$ that makes data most likely $\hat{\theta_{ml}} = \arg\max_\theta p_X(x;\theta)$-also applies when $x, theta$ are vectors or x is continuous
- compare to bayesian posterior:$p_{\Theta|X} = \frac{p_{X|\theta}p_\Theta}{p_X}$- interpretation is very different