11/29

LECTURE 21

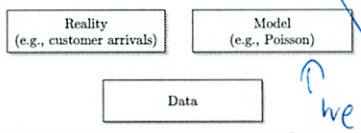
Last unit: statistics

- **Readings:** Sections 8.1-8.2

"It is the mark of truly educated people to be deeply moved by **statistics**." (Oscar Wilde)

+ inference

Samish thing



we concentrated on this part

models only useful if related to reality

- Design & interpretation of experiments
  - polling, medical/pharmaceutical trials...
- Netflix competition
- Finance

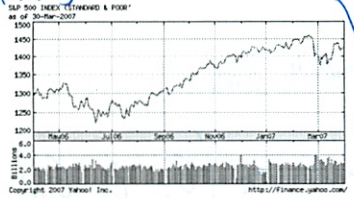
Collect data to validate model

delicate affair

manipulate data in sound matter

So data speaks for itself

movies try to predict rating of movie have not seen



extrapolate how stock is doing

"a technical stock analyst"

People

- Signal processing
  - Tracking, detection, speaker identification,...

How use this table

- Sophesticated probablistic model

Measure something, trying to figure out what is happening

best fit to the data

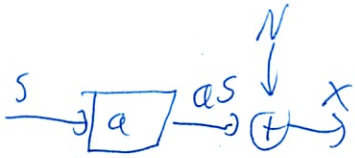
Types of Inference models/approaches

Main methodology

- Model building versus inferring unknown variables. E.g., assume $X = aS + W$
  - Model building: know "signal" S, observe X, infer a
  - Estimation in the presence of noise: know a, observe X, estimate S.
- **Hypothesis testing:** unknown takes one of few possible values; aim at small probability of incorrect decision - radar example
- **Estimation:** aim at a small estimation error

type of interpretation

1. Have data, build model →
2. Have model, get observations of outcomes X



attenuation need coefficient

Mathematically the two problems are the same

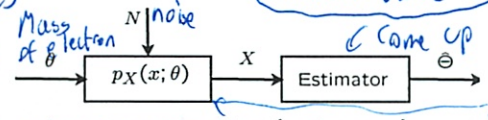
2 fundamental methodologies

How do we model the thing we do not know

- **Classical statistics:**

Some # - guess of unknown value

come up w/ estimate



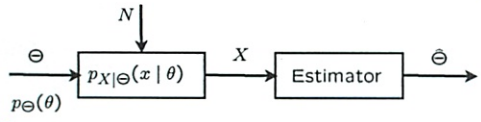
θ: unknown parameter (not a r.v.)
- E.g., θ = mass of electron

prob dist. Not conditional prob

Θ is a constant - but we don't know it

closer to what we were doing in class

- **Bayesian:** Use priors & Bayes rule



Collect data

improve knowledge of theta

) will focus on this for next few lectures

know theta is about here - RV

past accumulated beliefs of Θ

Bayesian inference: Use Bayes rule

- **Hypothesis testing**
  - discrete data

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

  - continuous data

$$p_{\Theta|X}(\theta \mid x) = \frac{p_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

- **Estimation;** continuous data

$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x \mid \theta)}{f_X(x)}$$

$$Z_t = \Theta_0 + t\Theta_1 + t^2\Theta_2$$
$$X_t = Z_t + W_t, \qquad t = 1, 2, \ldots, n$$

Bayes rule gives:

$$f_{\Theta_0,\Theta_1,\Theta_2|X_1,\ldots,X_n}(\theta_0, \theta_1, \theta_2 \mid x_1, \ldots, x_n)$$

Estimation with discrete data

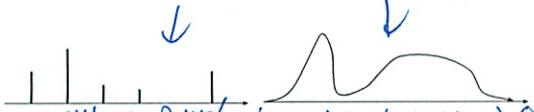
$$f_{\Theta|X}(\theta \mid x) = \frac{f_\Theta(\theta)\, p_{X|\Theta}(x \mid \theta)}{p_X(x)}$$

$$p_X(x) = \int f_\Theta(\theta) p_{X|\Theta}(x \mid \theta)\, d\theta$$

- **Example:**
  - Coin with unknown parameter θ
  - Observe X heads in n tosses
- What is the Bayesian approach?
  - Want to find $f_{\Theta|X}(\theta \mid x)$
  - Assume a prior on Θ (e.g., uniform)

Output of Bayesian Inference

- Posterior distribution:
  - pmf $p_{\Theta|X}(\cdot \mid x)$ or pdf $f_{\Theta|X}(\cdot \mid x)$



conditional PMF(θ) | conditional densities (θ)

- If interested in a single answer:
  - Maximum a posteriori probability (MAP):
    - $p_{\Theta|X}(\theta^* \mid x) = \max_\theta p_{\Theta|X}(\theta \mid x)$ minimizes probability of error; often used in hypothesis testing
    - $f_{\Theta|X}(\theta^* \mid x) = \max_\theta f_{\Theta|X}(\theta \mid x)$
  - Conditional expectation:

$$\mathbf{E}[\Theta \mid X = x] = \int \theta f_{\Theta|X}(\theta \mid x)\, d\theta$$

  - Single answers can be misleading!

will produce answer

want prob of making mistake is as small as possible

highest bar → call it θ^*

- the sum of the height of other bars

for continuous - how recievers in communication systems work

maximum density point (mode)

or look at median

or look at expected value / mean

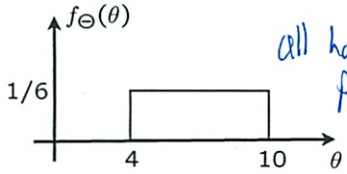
3 possible choices!

- all reasonable + plausible

- need to compare different ways of producing estimates

Least Mean Squares Estimation

- Estimation in the absence of information



all have is prior for θ

- find estimate c, to:

$$\text{minimize } \mathbf{E}\left[(\Theta - c)^2\right]$$

c estimate came up w/

- Optimal estimate: $c = \mathbf{E}[\Theta]$
- Optimal mean squared error:

$$\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta])^2\right] = \text{Var}(\Theta)$$

w/o any data!

most popular = mean squares

$\theta \rightarrow$ [model] $\xrightarrow{X}$ [estimator] $\rightarrow \hat{\theta}$

model: given

estimator: design

↳ function of data to produce an estimate $\hat{\theta} = g(x)$

$\theta - \hat{\theta}$ = error

look at square, b/c don't like errors

$(\theta - \hat{\theta})^2$ = penalty

- everything here is random!
- so keep small $E\left[(\theta - \hat{\theta})^2\right]$

Finding

$E[(\theta - c)^2] = E[\theta^2] - 2E[\theta]c + c^2$

$\frac{d}{dc} = 0$! $\quad -E[\theta] + c = 0$

$E[\theta] = c$

- this is what we did in 6.01!)

LMS Estimation of Θ based on X

before just prior dist $P_\Theta(\theta)$

- Two r.v.'s Θ, X
- we observe that $X = x$ — after observing $P_{\Theta|X}(\theta|x)$
  - new universe: condition on $X = x$
- $\mathbf{E}\left[(\Theta - c)^2 \mid X = x\right]$ is minimized by
  $c =$ $E[\Theta|X=x]$ ← because we had an observation
- $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X = x])^2 \mid X = x\right]$
  $\leq \mathbf{E}[(\Theta - g(x))^2 \mid X = x]$ ← chop & thinking of conditionals as RVs
  ↓ when $X=x$
  - $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2 \mid X\right] \leq \mathbf{E}\left[(\Theta - g(X))^2 \mid X\right]$
    $E[$ $]$ $E[$ $]$ ← can do iterated expectations
  - $\mathbf{E}\left[(\Theta - \mathbf{E}[\Theta \mid X])^2\right] \leq \mathbf{E}\left[(\Theta - g(X))^2\right]$ $\forall g$ - g is estimator

is optimal

$\mathbf{E}[\Theta \mid X]$ minimizes $\mathbf{E}\left[(\Theta - g(X))^2\right]$ over all estimators $g(\cdot)$

↳ smallest possible over all estimators

estimate = # $\hat{\theta}$ $E[\theta|X=x]$

estimator = formula or algorithm $\hat{\theta}$

↳ conditional expectation $E[\theta|X]$

LMS Estimation w. several measurements

lots of data

- Unknown r.v. Θ
- Observe values of r.v.'s $X_1, \ldots, X_n$ — can still make same argument
- Best estimator: $\mathbf{E}[\Theta \mid X_1, \ldots, X_n]$ — sometimes hard to calculate
- Can be hard to compute/implement — lots of joint prob
  - involves multi-dimensional integrals, etc.

- sometimes may have to settle for less than this
- devil is in the details

11/29

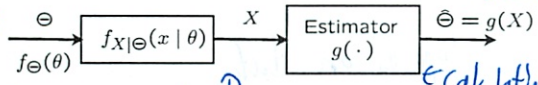
LECTURE 22

- **Readings:** pp. 225-226; Sections 8.3-8.4

Topics

2nd lecture

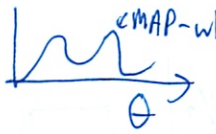
- (Bayesian) Least means squares (LMS) estimation
- (Bayesian) Linear LMS estimation



← get estimate of θ
-function of data
← calculation
Obtain measurement x
Unknown quantity - treat as RV

- MAP estimate: $\hat{\theta}_{\text{MAP}}$ maximizes $f_{\Theta|X}(\theta \mid x)$
- LMS estimation:
  - $\hat{\Theta} = \mathbf{E}[\Theta \mid X]$ minimizes $\mathbf{E}\big[(\Theta - g(X))^2\big]$ over all estimators $g(\cdot)$
  - for any x, $\hat{\theta} = \mathbf{E}[\Theta \mid X = x]$ minimizes $\mathbf{E}\big[(\Theta - \hat{\theta})^2 \mid X = x\big]$ over all estimates $\hat{\theta}$

Obtain Posterior dist



← MAP - where tallest
-most likely value

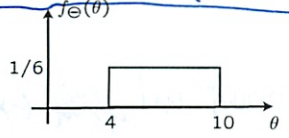
how far estimate is from

Sometimes good estimator
-What is good way of getting estimate

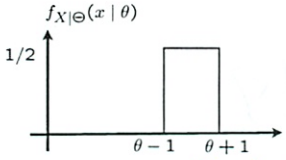
treat as # if give specific value of data

squared value
- simple solution
$E[\Theta \mid X]$

-recalculate in new universe → find mean → is best estimate



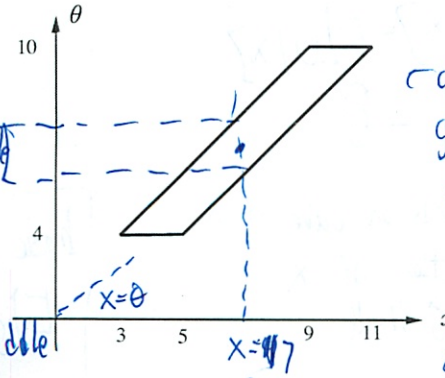
← when unknown variable



← get observation X

$X = \Theta + M$ ← noise, uniform [-1, 1]

↑ observation error (from 6.01 I think)



← apply Bayes rule
get joint
divide by marginal

$f_{\Theta|X} = \dfrac{f_{\Theta,X}}{f_X}$

$f_{\Theta,X} = f_\Theta \cdot f_{X|\Theta} = \dfrac{1}{12}$

← for possible values

Can see possible values of Θ
-Uniform dist
mean/midpoint is in the middle
ends up being

when told an X

← $g(X) = E[\Theta \mid X = x]$

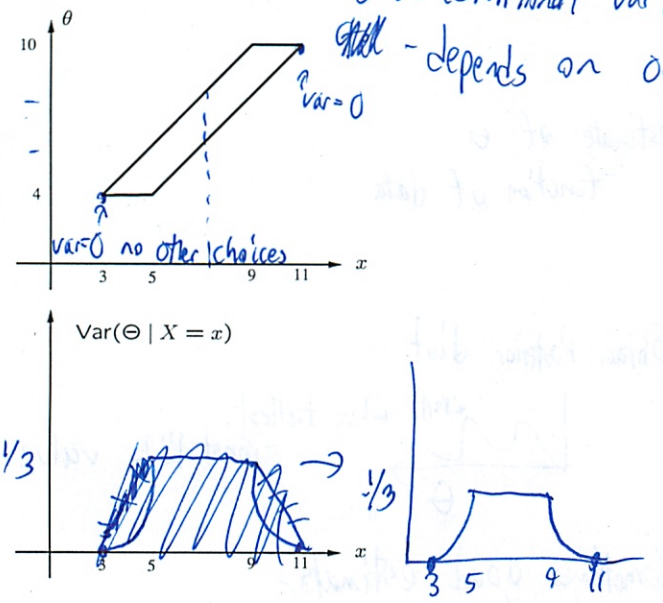
~~$f_{\Theta,X} = f_\Theta \cdot f_{X|\Theta}$~~
~~$= \frac{1}{12}$ for possible values of Θ, X possible~~

(nice - I like this!)

How good is your estimate?

Conditional mean squared error

- $\mathbf{E}[(\Theta - \mathbf{E}[\Theta \mid X])^2 \mid X = x]$ ← trying to minimize
  - same as $\mathrm{Var}(\Theta \mid X = x)$: variance of the conditional distribution of Θ ← just conditional var!
    - depends on observed X



Possible error
- over this uniform interval
$\frac{(b-a)^2}{12} = \frac{1}{3}$

depends on data
- Some data very useful
- Since small var

Var = estimation error

Some properties of LMS estimation

"Heavy lifting"
math properties of conditional expectations,

- Estimator: $\hat{\Theta} = \mathbf{E}[\Theta \mid X]$
- Estimation error: $\tilde{\Theta} = \hat{\Theta} - \Theta$
  - what is the avg value of estimation error
- $\mathbf{E}[\tilde{\Theta}] = 0$ $\quad \mathbf{E}[\tilde{\Theta} \mid X = x] = 0$
- $\mathbf{E}[\tilde{\Theta} h(X)] = 0$, for any function h
- $\mathrm{cov}(\tilde{\Theta}, \hat{\Theta}) = 0$
- Since $\Theta = \hat{\Theta} - \tilde{\Theta}$:
  $\mathrm{var}(\Theta) = \mathrm{var}(\hat{\Theta}) + \mathrm{var}(\tilde{\Theta})$

$$E[\tilde{\theta} \mid X] = E[\hat{\theta} - \theta \mid X]$$

linearity of expectation

$$= E[\hat{\theta} \mid X] - E[\theta \mid X]$$

$$= \hat{\theta} - \hat{\theta}$$

↑ depends on value of X
function of X
so $\hat{\theta}$ itself

$$= 0$$

basic principle
$E[g(X) \mid X] = g(X)$

$E[\tilde{\theta} \mid X = x] = 0$ still 0 when have specific x

law of iterated expectations
$E[\tilde{\theta}] = E[E[\tilde{\theta} \mid X]] = E[0] = 0$

So on avg error will be 0
- unbiased

See other sheet

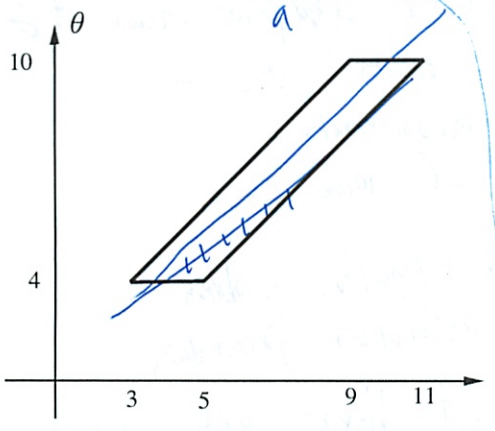
Linear LMS

Simplify a bit

- Consider estimators of Θ, of the form $\hat{\Theta} = aX + b$
- Minimize $\mathbf{E}\left[(\Theta - aX - b)^2\right] = h(a,b)$ (squared error)
- Best choice of a, b; best linear estimator:

$$\hat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(X,\Theta)}{\mathrm{var}(X)}(X - \mathbf{E}[X])$$

($\frac{\mathrm{Cov}(X,\Theta)}{\mathrm{var}(X)}$ = a)



- if too difficult to calc. conditional expectation
- instead of getting optimal estimator
- just get a single line
- easier to calculate/estimate
- losing optimality

Find constants a, b, so mean squared line is least

Expand as function of a, b

$\frac{\partial h}{\partial a} = 0 = \frac{\partial h}{\partial b}$ - will be quadratic in a, b
- will get linear eq
- solve

Optimal form is this

Linear LMS properties

if no data, a disappears, or if they are ind.
if have data, add correction information, new info

$$\hat{\Theta}_L = \mathbf{E}[\Theta] + \frac{\mathrm{Cov}(X,\Theta)}{\mathrm{var}(X)}(X - \mathbf{E}[X])$$

(single RV)

$$\mathbf{E}[(\hat{\Theta}_L - \Theta)^2] = (1 - \rho^2)\sigma_\Theta^2$$

mean squared estimation error

↳ original value of Θ reduced by a certain factor
- if ρ close to 1, error becomes small
- exploiting correlations b/w data + unknown quant
- Can reduce error if data somewhat correlated

Linear LMS with multiple data

(multiple observations)

- Consider estimators of the form:

$$\hat{\Theta} = a_1 X_1 + \cdots + a_n X_n + b$$

(linear to keep it simple)

- Find best choices of $a_1, \ldots, a_n, b$
- Minimize:

$$\mathbf{E}[(a_1 X_1 + \cdots + a_n X_n + b - \Theta)^2]$$

Choice of coeffects to get smallest possible error

↳ lots of calculations
do w/ computer

- Set derivatives to zero
linear system in b and the a_i
- Only means, variances, covariances matter

↑ - also cov of cross terms

- detailed stats don't matter (dists)
- only things that matter is these

The cleanest linear LMS example

Can be tedious

$X_i = \Theta + W_i, \qquad \Theta, W_1, \ldots, W_n$ independent
$\Theta \sim \mu,\ \sigma_0^2 \qquad W_i \sim 0,\ \sigma_i^2$

$$\hat{\Theta}_L = \frac{\mu/\sigma_0^2 + \sum_{i=1}^{n} X_i/\sigma_i^2}{\sum_{i=0}^{n} 1/\sigma_i^2}$$

(weighted average of $\mu, X_1, \ldots, X_n$)

- If all normal, $\hat{\Theta}_L = \mathbf{E}[\Theta \mid X_1, \ldots, X_n]$

& less tedious, get clean formula

Observe several noisy measurements
Usually avg measurements
- which is roughly ~~our~~ optimal
- but want to include ~~fixed~~ prior mean of Θ
- but not all measurements are =
  - must weight measurements
  - weights inv of noise vars

Choosing X_i in linear LMS

- $\mathbf{E}[\Theta \mid X]$ is the same as $\mathbf{E}[\Theta \mid X^3]$
- Linear LMS is different:
  - $\hat{\Theta} = aX + b$ versus $\hat{\Theta} = aX^3 + b$
  - Also consider $\hat{\Theta} = a_1X + a_2X^2 + a_3X^3 + b$

& same info, so same estimation problem
- won't change estimation procedure
- Unless looking at linear estimators
- need to look at different estimators
- do what you think will fit the data better

$\begin{bmatrix} X \\ X^2 \\ X^3 \end{bmatrix}$ different form again

- art in what you call the data in problem
  - LMS does not matters
  - matters in linear

Some orientation

Review

- **Standard examples:**
  - X_i uniform on $[0, \theta]$; uniform prior on θ
  - X_i Bernoulli(p); uniform (or Beta) prior on p
  - X_i normal with mean θ, known variance σ^2; normal prior on θ; $X_i = \Theta + W_i$
- **Estimation methods:**
  - MAP
  - MSE
  - Linear MSE

$$E[\tilde{\theta} h(x) | X] = E[\hat{\theta} h(x) | X] - E[\theta h(x) | X]$$

if x known, just the values

make outside since quantity known — nothing random about it

$$= \hat{\theta} h(x) - h(x) E[\theta | X]$$

$$= \hat{\theta} h(x) - h(x) \hat{\theta}$$

$$= 0$$

$E[\tilde{\theta} h(x)]$

↑ uncorrelated w/ the error

no function of data that gives us useful info about the error

could use to ↓ error further

but if optimal, can not ↓ error further

Use this to look at cov

$$\text{Cov}(\tilde{\theta}, h(x)) = E[\tilde{\theta} h(x)] - E[\tilde{\theta}] E[h(x)]$$

$$= 0 - 0 \; E[h(x)]$$

$$= 0$$

(2)

Apply to special case where $h(x) = \hat{\theta}$

$\hat{\theta}$ is just function of x

$$Cov(\tilde{\theta}, \hat{\theta}) = 0$$

* error and estimate are uncorrelated!

- Since uncorrelated, add vars

- So having big var in estimate is good

  $\hat{\theta}$ big → $\tilde{\theta}$ affected by the data

  ↳ when the data is informative

  helps make $\tilde{\theta}$ small → so small error

- Can prove w/ law of total variance

  = var(estimate) + var(error)

More Chap 4 Reading 11/29

5.4 Central Limit Theorm

- according to WLLN distribution of sample mean increasingly concentrated in the near vicinity of true mean μ
- Var tends to 0
- so $S_n = X_1 + \dots + X_n = nM_n \to \infty$ and distribution does not converge to any thing meaningful
- do this scaling $Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}}$

$$E[Z_n] = \frac{E[X_1 + \dots + X_n] - n\mu}{\sigma\sqrt{n}} = 0$$

$$\text{Var}(Z_n) = \frac{\text{var}(X_1 + \dots + X_n)}{\sigma^2 n} = \frac{\text{Var}(X_1) + \dots + \text{var}(X_n)}{\sigma^2 n} = \frac{n\sigma^2}{n\sigma^2} = 1$$

Central Limit Theorm

$$Z_n = \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \qquad \mu = \text{mean}, \quad \text{Var} = \sigma^2$$

CDF Z_n converges to st normal for every z

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx \to \lim_{n\to\infty} P(Z_n \le z) = \Phi(z)$$

(2)

Very general - only requires ind, finite mean, var

Shows sum of large # of ind. RVs is ~ normal

applies to many situations (been proven empirically)

Also makes hard calculations simpler

↳ allows us to calculate probabilities related to Z_n

↳ like treating S_n as normal

w/ mean $n\mu$

var $n\sigma^2$

Normal Approx based on Central Limit Theorm

$S_n = X_1 + \cdots + X_n$ X_i iid RVs mean μ var σ^2

If n is large $P(S_n \leq c)$ ~ normal by:

1. Calculate mean $n\mu$ var $n\sigma^2$ of S_n
2. Calculate the normalized value $z = \frac{(c - n\mu)}{\sigma\sqrt{n}}$
3. Use approx $P(S_n \leq c) \approx \Phi(z)$

↳ from st. normal tables

(3)

example

have 100 packages that weigh b/w 5 → 50 lbs

P(weight < 3800 lbs)

easier to do w/ central limit

$S_{100} = X_1 + \dots + X_{100}$

↳ $\mu = \frac{5+50}{2} = 27.5$

? where get this? ch dist of ~~asdfa~~ uniform RV

$\sigma^2 = \frac{(50-5)^2}{12}$

Then normalize:

$$z = \frac{3000 - 100 \cdot 27.5}{\sqrt{168.75 \cdot 100}} = \frac{250}{129.9} = 1.92$$

Plug into st. normal table

$P(S_{100} \leq 3000) \approx \Phi(1.92) = .9726$

So $P(S_{100} > 3000) = 1 - P(S_{100} \leq 3000)$

$= .0274$

(4)

If var X_i is unknown can use upper band

- example Still polling

$$M_n = \frac{X_1 + \cdots + X_n}{n}$$

$M_n \to$ mean p var $\frac{p(1-p)}{n}$

(? are they sure not X_i) So M_n is approx normal

Interested in $P(|M_n - p| \geq \epsilon)$

- prob that polling error is larger than desired accuracy ϵ

$$P(|M_n - p| \geq \epsilon) \approx 2\, P(M_n - p \geq \epsilon)$$

Symmetry of normal PDF around mean

Var $\frac{p(1-p)}{n}$ of $M_n - p$ depends on p so unknown

So assume largest possible var $\frac{1}{4n}$ (erased this)

which corresponds to $p = \frac{1}{2}$

(5)

Now use this ~~Standardized~~ to find standardized value

$$z = \frac{\varepsilon}{1/(2\sqrt{n})}$$

$$P(M_n - p \geq \varepsilon) \leq 1 - \phi(z)$$

$$= 1 - \phi(2\varepsilon\sqrt{n})$$

So if $n = 100$ $\varepsilon = .1$

$$P(M_{100} - p| \geq .1) \approx 2P(M_n - p \geq .1)$$

$$\leq 2 - 2\phi(2 \cdot .1 \cdot \sqrt{100})$$

$$= 2 - 2\phi(2)$$

$$= 2 - 2 \cdot .977$$

$$= .046$$

Much smaller / more accurate than Chebshev

Normal approx is increasingly accurate as $n \to \infty$

But usually we have a finite n

No simple ans when n is big enough

(6)

De Moivre-Laplace Approx to Binomial

- (: did we do this in lecture?)
- Binomial RV S_n w/ param n, p
  - Sum of n ind. Bernoulli RV $X_1, \ldots, X_n$ w/ common param p

$$S_n = X_1 + \cdots + X_n$$

$$\mu = E[X_i] = p \qquad \sigma^2 = \sqrt{\text{var}(X_i)} = \sqrt{p(1-p)}$$

Now want prob $k \le S_n \le \ell$

$$\frac{k-np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\ell - np}{\sqrt{np(1-p)}}$$

$\uparrow$ approx st. normal

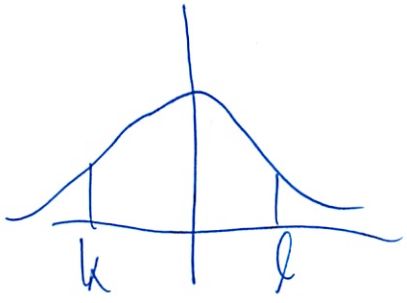
$$P(k \le S_n \le \ell) = P\left(\frac{k-np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\ell - np}{\sqrt{np(1-p)}}\right)$$

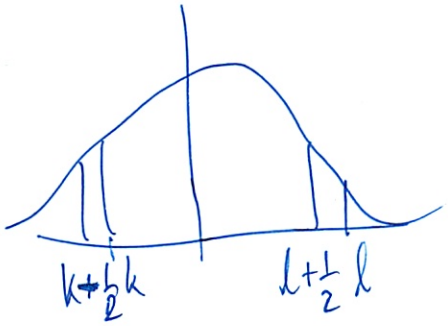
$$\approx \Phi\left(\frac{\ell - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$

$$\approx \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)$$

(7)

- got more accurate by replacing $k \to k - \frac{1}{2}$
  $\ell \to \ell + \frac{1}{2}$





; so how does that help?

skipping

5.5 Strong Law of Large

- similar to weak law
  - also about convergence of sample mean to true mean
- but its a different type of convergence

$X_1, X_2, \ldots$ are iid RVs w/ mean μ

$M_n = \frac{(X_1 + \cdots + X_n)}{n}$ converges to μ

w/ prob 1

$P\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) = 1$

States that all of prob concentrated on a particular subset of sample set

(kinda confused on what it says)

Convergence w/ Prob 1

Let $Y_1, Y_2, \ldots$ be seq of RVs (not necessarily ind)

C = real #

Say Y_n converges to C w/ prob 1 (surely) if

$$P\left(\lim_{n \to \infty} Y_n = C\right) = 1$$

All prob concentrated on seq that converge to C

- other seq unlikely

Convergence w/ prob 1 implies Conv. in prob

Recitation 22
November 30, 2010

Examples 8.2, 8.7, 8.12, and 8.15 in the textbook

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown and is modeled as the value of a random variable Θ, uniformly distributed between zero and one hour.

(a) Assuming that Juliet was late by an amount x on their first date, how should Romeo use this information to update the distribution of Θ?

(b) How should Romeo update the distribution of Θ if he observes that Juliet is late by $x_1, \ldots, x_n$ on the first n dates? Assume that Juliet is late by a random amount $X_1, \ldots, X_n$ on the first n dates where, given θ, $X_1, \ldots, X_n$ are uniformly distributed between zero and θ and are conditionally independent.

(c) Find the MAP estimate of Θ based on the observation $X = x$.

(d) Find the LMS estimate of Θ based on the observation $X = x$.

(e) Calculate the conditional mean squared error for the MAP and the LMS estimates. Compare your results.

(f) Derive the linear LMS estimator of Θ based on X.

(g) Calculate the conditional mean squared error for the linear LMS estimate. Compare your answer to the results of part (e).

Recitation 22

11/30

- graded p-set 8+9 on shelves outside TA office

- most important Bayes problem in recitation 2
  - the disease one
  - (also in 6.01)
  - archtypal example
  - update beliefs when have info
- just more complex in this chap
- parameter θ
  - random
    - known prob density = prior
- $f_{X|\Theta}(x|\theta)$ = math description of how obs generated
  - likelyhood function
  - obs model
  - generative model

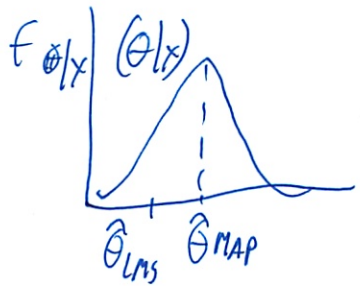
) aka
- $f_{\Theta|X}(\theta|x)$ $\quad \Theta$ = capital θ
  - want to get via Bayes rule
  - posterior

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

(2)

Is everything we know
- sometimes too much

Want to boil down to 1 #
- maximum point on posterior = MAP
- mean of posterior = LMS
  - $E[\Theta | X = x]$
  - Minimizes $E[(\Theta - \hat{\theta})^2 | X = x]$



#1. Romeo and Juliet

- don't know θ
- as go on dates, she know late she is, increase knowledge of θ

$f_\Theta(\theta) = \begin{cases} 1 & 0 \le \theta \le p \\ 0 & \text{otherwise} \end{cases}$ initial dist / prior

Obs model

$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{\theta} & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$

a) $F_{\Theta|X}(\theta|x)$ ← saw one little x since went on 1 date

$$= \begin{cases} \frac{1 \cdot \frac{1}{\theta}}{\int_x^1 \frac{1}{\theta} d\theta} & \text{normalize} \quad 0 \le x \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases}$$

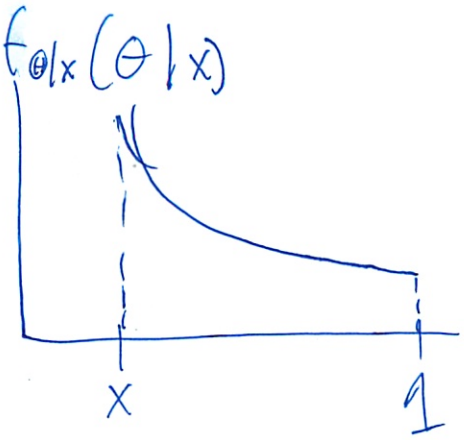
evaluate

$$= \begin{cases} \frac{1}{-\theta \log x} & 0 \le x \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$= \begin{cases} \frac{1}{\theta \log(\frac{1}{x})} & \cdots \\ 0 & \cdots \end{cases}$$

another way to write

(3)



height $\propto \frac{1}{\theta}$ shape

- other just normalization so $\int$ to 1

Romeo and Juliet late on 1st date

b) Now go on n dates

$$f_{\theta|X_1,X_2,\ldots,X_n}(\theta \mid X_1, X_2, \ldots, X_n)$$

$$= \frac{f_\theta(\theta)\, f_{X_1,X_2,\ldots X_n|\theta}(X_1, X_2, \ldots, X_n|\theta)}{\int \cdots d\theta}$$

Write as product of marginal conditional dist.

$$= \frac{f_\theta(\theta) f_{X_1|\theta}(X_1|\theta) \cdots f_{X_n|\theta}(X_n|\theta)}{\int \cdots d\theta}$$

$$= \begin{cases} \dfrac{\cdots}{\int \cdots d\theta} & \cdots \\ 0 & \text{otherwise} \end{cases}$$

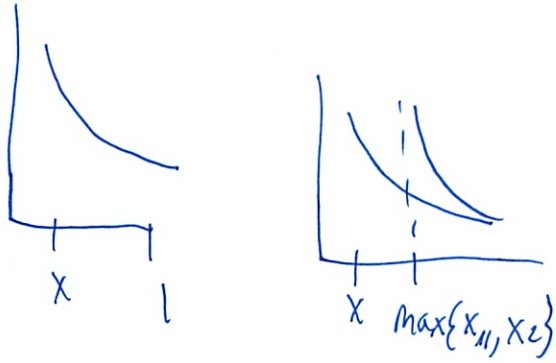
When is the first factor between X_1 and 1?

?? confused

(4)

$$= \begin{cases} \dfrac{\frac{1}{\theta^n}}{\int_{\max\{x_i\}}^{1} \frac{1}{\theta^n} d\theta} & \max\{x_i\} \le \theta \le 1 \\ 0 & \text{otherwise} \end{cases}$$

For larger n $\llcorner$ becomes less flat / steeper



c) Want a single # estimate
- here from 1 observation

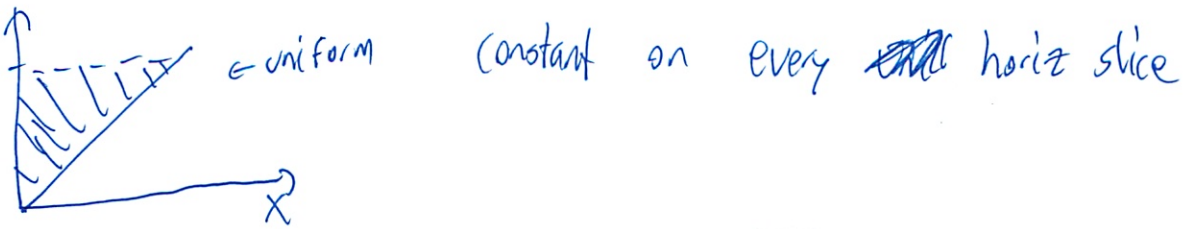
$\hat{\theta}_{MAP}$ = value of θ maximum of posterior

notice that is decreasing function of θ

So is x

$\hat{\theta}_{MAP} = x$

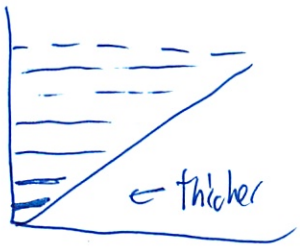
extra) Sketch joint dist



(5)

So need taller slices for tall values of θ
- for slices to ~~1~~ 5 to 1

Observe X taking slice in 1 direction
- taller for small values of θ



Map just takes left end of distribution
But picking 1 end might not be best estimate
Minimize mean of square of error = LMS
Will be to left of mid point

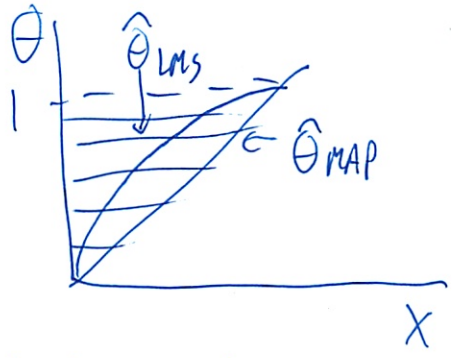
d) $\hat{\theta}_{LMS} = E[\Theta | X = x]$

$= \int \theta \, f_{\Theta|X}(\theta|x) \, d\theta$
(posterior density)

$= \int_x^1 \theta \frac{1}{\theta \log(\frac{1}{x})} \, d\theta$

$= \frac{1-x}{\log(\frac{1}{x})}$

Then on our plot it would be
- at each point



e) Most tedious

Find conditional mean squared error for MAP, LMS

$$E\left[(\Theta-\hat{\theta})^2 \mid X=x\right] =$$

* Can make computations that apply in both cases

$$= \int (\theta-\hat{\theta})^2 f_{\Theta|X}(\theta \mid x)\, d\theta$$

$$= \int_x^1 (\theta-\hat{\theta})^2 \frac{1}{\theta \log\left(\frac{1}{x}\right)} d\theta$$

tedious, not hard calculation

$$= \int_x^1 \left(\theta^2 - 2\theta\hat{\theta} + \hat{\theta}^2\right) \frac{1}{\theta \log\left(\frac{1}{x}\right)} d\theta$$

$$= \hat{\theta}^2 - \hat{\theta}\frac{2(1-x)}{\log\left(\frac{1}{x}\right)} + \frac{1-x^2}{2\log\left(\frac{1}{x}\right)}$$

for any deterministic function $\hat{\theta}$ - can put in LMS, MAP

(7)

So if put in $\hat{\Theta}_{MAP}$ → Conditional Mean Squared Error (MSE)

$$X^2 + \frac{3x^2 - 4x + 1}{2\log(\frac{1}{x})}$$

$\hat{\Theta}_{LMS}$

$$\frac{1-x^2}{2\log(\frac{1}{x})} - \left(\frac{1-x}{\log(\frac{1}{x})}\right)^2$$

No insights to draw

Compare the two

LMS must be smaller than MAP by definition

For all possible value of X

Plots in textbook

e) Want to find the LLMS estimator

- minimizes LMS among all estimators that are functions of x

$$\hat{\Theta}_{LLMS} = E[\Theta] + \frac{Cov(X,\Theta)}{Var(X)}(X - E[X])$$

- defined in lecture
- "plausably good"
- will not reprove

(8)

How to compute the 4 constants?

$E[\Theta] = \frac{1}{2}$

$E[X]$ = have avoided calculating dist of X
Can avoid calculating it
Conditioning on Θ makes X simple
So do iterated expectations

$= E[E[X \mid \Theta]]$

↑ conditioning on RV subtile constant

$= E[E[X \mid \Theta \neq \theta]]$

↑ calculate for little θ

E of $\theta = \frac{\theta}{2}$

$= E[\Theta/2]$ then back to capital letters

$= \frac{1}{4}$

$\mathrm{Var}(X)$ = use law of total var

$= E[\mathrm{var}(X \mid \Theta)] + \mathrm{var}(E[X \mid \Theta])$

$= E[\frac{1}{12}\Theta^2] + \mathrm{var}(\frac{\Theta}{2})$

var + E]^2

$= \frac{1}{12}\left(\frac{1}{12} + \left(\frac{1}{2}\right)^2\right) + \frac{1}{4} \cdot \frac{1}{12}$

trying to do w/ properties of uniform dist

(9)

$$= \frac{1}{12} \cdot \frac{1}{3} \cdot \frac{1}{4} \cdot \frac{1}{12} = \frac{7}{144}$$

$$Cov(X, \Theta) = E[X\Theta] - E[X]E[\Theta]$$

$$= ? - \frac{1}{4} \cdot \frac{1}{2}$$

$?$

$$E[X\Theta] = E[E[X\Theta \mid \Theta]]$$

↑ condition on Θ - easier

- find for little θ, then replace w/ Θ

$$E[X\Theta \mid \Theta = \theta]$$

$$= E[\theta X \mid \Theta = \theta]$$

$$= \theta E[X \mid \Theta = \theta]$$

$$= \theta \cdot \frac{\theta}{2}$$

$$= E\left[\frac{1}{2}\Theta^2\right]$$

$$= \frac{1}{2} \cdot \frac{1}{3}$$

↑ just did

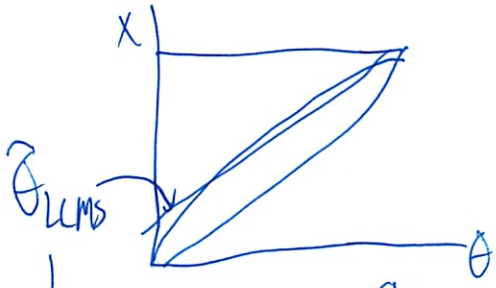
$$= \frac{1}{6}$$

$$Cov(X, \Theta) = \frac{1}{24}$$

(10)

$$\hat{\theta}_{LLMS} = E[\theta] + \frac{Cov(X,\theta)}{Var(X)}(X - E[X])$$

$$= \frac{1}{2} + \frac{6}{7}\left(X - \frac{1}{4}\right)$$



↳ like the best fit of the $\hat{\theta}_{LMS}$ line

(Very cute how can draw this diagram - I want to become good at this)

Bayesian Statistical Inference Chap 8 Reading

11/[illegible]
36

- process of extracting info about an unknown variable or an unknown model from available data
- 2 main approaches: Bayesian + classical
- main categories of inference problems: parameter estimation, hypothesis testing, significance testing
- important methodologies: MAP, LMS, max likelihood, regression, likelihood ratio tests

Probability vs Stats

- probability is a self-contained math subject
  - assume a fully specified prob model
  - use math to quantify consequences of model
  - every qu has an answer
  - model taken for granted
- Statistics more of an art form
  - may be several reasonable methods → several answers
  - no 1 way to select right answer
  - so require certain desirable principals
    - performance
    - past experience
    - common sense
    - community

(2)

Bayesian vs Classical Stats

Bayesian – unknowns treated as RVs

Classical/frequentist – deterministic quantity, only unknown

Bayesian tries to move stats back to probability

* dealing with multiple possible candidate probabilistic models *

Is (which are these again? $\hat{\theta}_{MAP}$, etc?)

Each one thinks their methods is better
(Interesting discussion and comparison)

Model vs Variable Inference

Model inference – real phenomenon or process

want to construct/validate a model

have available data

to predict about the future or infer some underlying cause

Variable inference – estimate value of 1 or more unknown values w/ noisy info

- Some overlap b/w the models
- may do both

A Rough Classification of Stat. Inference Problems

Estimation - model fully specified

- except for unknown, possibly multidimensional parameter θ
  - want to estimate close to its exact value
  - - RV (Bayesian)
    - unknown constant (classical)
- noisy transmition → want to estimate a (model of channel)
- using polling data, estimate fraction of people that prefer one thing to the other
- estimate mean, var of movement of a stock

Binary hypothesis testing - start w/ 2 hyp. decide which is True

- use knowledge of a, X_i to decide if $s_i = 0$ or 1
  - a: noise model
  - X_i: recieved bit
  - s_i: sent bit
- with a noisy pic → is there a person in there or not?
- decide which of 2 medical treatments is best

M-ary hypothesis testing

- finite # m of hypothesis

Perfomance of a model judged by prob that *it* makes an erroneous decision

This chap focuses on estimation

Ignoring more advanced topics like nonparametric $Y = g(X) + W$

- g: unknown function (can't be described by fixed # parameters)
- W: noise

Principal Bayesian Inference Methods

Maximum a posteriori prob (MAP)

- the max on the posterior dist

Least Mean Squares (LMS) estimation

Select function/estimator that minimizes mean squared error b/w parameter and its estimate

Linear least Mean Squares (LLMS)

- linear function of data, like above
- simpler, but a little less accurate

8.1 Bayesian Inference + Posterior Distribution

Θ = capital θ = Unknown quantity of interest
- ↳ RV
- ↳ or Finite # of RVs
- ↳ can represent physical quantities

$X = (X_1, X_2, \ldots, X_n)$ = Observations
Observation = measurements / aka vector

Know

joint dist Θ, X

OR

prior dist p_Θ / f_Θ

Conditional dist $p_{X|\Theta} / f_{X|\Theta}$

Once X has been observed, get posterior dist

$p_{\Theta|X}(\theta|x) / f_{\Theta|X}(\theta|x)$

↑ get via bayes rule

(4)

Remember are 4 versions of Bayes rule for various combos
Continuous + discrete

Example - did today in recitation
(why math so complex - what's going on again?)

- other examples complex

recursive inference - adding X_{n+1} observations does not change posterior dist

- Can also have multiparameter problems
  - won't have closed form solution however
  - sometimes you will be just focused on 1 of the components

---

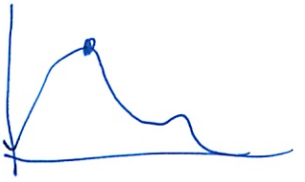
8.2 MAP

(how can there be a whole chap on this - so easy)

$\hat{\theta}_{MAP} = \arg\max_{\theta} P_{\theta|x}(\theta|x)$ discrete

" " " $f_{\theta|x}(\theta|x)$ continuous

Just the max



- maximizes the prob of a correct decision for any given value given X
- since tallest = most likley
- So only need to calculate numerator of posterior dist (shortcut)

Point Estimation

~~given the observed value x of X, posterior ca~~

certain quantities that summarize posterior

- single # value of Θ

$\hat{\Theta}$ = estimate

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

realized value $= g(x)$ when X takes value x.

depends on the observation

MAP is a estimator

Conditional Expectation estimator $\hat{\Theta} = E[\Theta | X = x]$

↳ Aka LMS estimator

(why is all this here - seems out of place)

(8)

if Posterior distribution Θ is symmetric around its conditional mean and is unimodal (single max), Max is at mean

MAP = LMS

(duh)

if Θ continuous $\hat{\theta}_{MAP}$ can be found analytically

- set to deriv to 0
- otherwise # search

Both MAP and LMS have no guarentees on accuracy

Hypothesis Testing

Θ takes one of m values $\theta_1, \theta_2, \ldots \theta_m$

- m = 2 when binary

event $\{\Theta = \theta_i\}$ = ith hypothesis = H_i

After obs x, calculate posterior probabilities

$$P(\Theta = \theta_i \mid X = x) = p_{\Theta|X}(\theta_i \mid x) \text{ for each } i$$

Select hyp whose prob is highest (MAP)

Can calculate prob of a corresponding decision

$g_{MAP}(x)$ = Observed hyp

$$P(\Theta = g_{MAP}(x) \mid X = x)$$

(9)

Overall prob of correct decisions is

$$P(\Theta = g_{MAP}(X)) = \sum_i P(\Theta = \theta_i, X \in S_i)$$

Corresponing prob of error is

$$\sum_i P(\Theta \neq \theta_i, X \in S_i)$$

(skipping examples for now)

8.3 Bayesian ~~Least~~ Mean Squares Estimation (LMS)

aka the conditional expectations estimator

Simpler problem

Estimate Θ w/ constant $\hat{\theta}$ w/o observation X

estimation error $= \hat{\theta} - \Theta$ = RV

$E[(\Theta - \hat{\theta})^2]$ is a #

- depends on $\hat{\theta}$
- can be minimized over $\hat{\theta}$

Best possible estimate $\rightarrow \hat{\theta} = E[\Theta]$

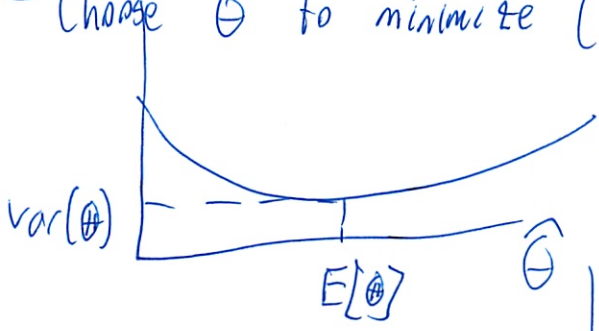
So for any estimate $\hat{\theta}$ have

$$E[(\Theta - \hat{\theta})^2] = var(\Theta - \hat{\theta}) + (E[\Theta - \hat{\theta}])^2$$

$$= Var(\Theta) + (E[\Theta] - \hat{\theta})^2$$

(10)

- Uses $E[Z^2] = Var(Z) + (E[Z])^2$
- Choose $\hat{\theta}$ to minimize $(E[\theta - \hat{\theta}])^2$



$$\hat{\theta} = E[\theta]$$

Now use observation X to estimate θ, so minimize MSE

Once know $X = x$ - now in new universe conditioned on $X = x$

So $E[\theta | X = x]$ minimizes conditional MSE $E[(\theta - \hat{\theta})^2 | X = x]$ over all constants $\hat{\theta}$

Unconditional w/ estimator

$$E[(\theta - g(X))^2]$$

$E[\theta | X]$ is estimator so MSE minimized when $g(X) = E[\theta | X]$

Example from lecture

Some properties of estimation error

$\hat{\Theta} = E[\Theta | X]$ $\quad \tilde{\Theta} = \hat{\Theta} - \Theta$

$\tilde{\Theta}$: estimation error; $\hat{\Theta}$: estimate

$\tilde{\Theta}$ is unbiased $\Leftrightarrow E[\tilde{\Theta}] = 0$

$E[\tilde{\Theta} | X = x] = 0$ for all x

$\tilde{\Theta}$ is uncorrelated w/ $\hat{\Theta}$

$Cov(\hat{\Theta}, \tilde{\Theta}) = 0$

$Var(\Theta)$ decomposes $Var(\Theta) = Var(\hat{\Theta}) + Var(\tilde{\Theta})$

↑ if 0 then info X is uninformative

Case of Multiple Observations and Multiple Parameters

if X is not just single RV, but vector of RVs

MSE estimation error minimized when use $E[\Theta | X_1, \ldots, X_n)$ as estimator

$$E[(\Theta - E[\Theta | X_1, \ldots, X_n])^2] \leq E[(\Theta - g(X_1, \ldots, X_n))^2]$$

for all estimators $g(X_1, \ldots, X_n)$

but hard to implement
- need a joint PDF $f_{\theta, X_1, X_2, \ldots, X_n}$
- expected function hard to calculate even if have joint PDF

If want to estimate multiple parameters $\theta_1, \ldots \theta_m$

$$E[(\theta_1 - \hat{\theta}_1)^2] + \ldots + E[(\theta_m - \hat{\theta}_m)^2]$$

and minimize over all estimators $\hat{\theta}_1, \ldots \hat{\theta}_m$

So decouple estimation problem

8.4 Bayesian Linear Least Mean Squares Estimation (LLMS)

- restricted class of estimators to those that are linear functions of observations
- might have higher mean squared error
- but much easier to calculate

$$\hat{\theta} = a_1 X_1 + \ldots + a_n X_n + b$$

↑ scalers

Corresponding MSE

$$E[(\theta - a_1 X_1 - \ldots - a_n X_n - b)^2]$$

↑ Chose these to minimize

but first where $n = 1$

(13)

Single Observation

$E[(\Theta - aX - b)^2]$ is associated w/ a linear estimator $aX+b$ of Θ

- if a already chosen → how chose b

$$b = E[\Theta - aX] = E[\Theta] - a\,E[X]$$

- with this b, now minimize w/ respect to a

$$E[(\Theta - aX - E[\Theta] + a\,E[X])^2]$$

- ~~[illegible]~~
- rewrite as ~~[illegible]~~

$$\mathrm{Var}(\Theta - aX) = \sigma_\Theta^2 + a^2\sigma_X^2 + 2\mathrm{cov}(\Theta, -aX)$$

$$= \sigma_\Theta^2 + a^2\sigma_X^2 - 2a\,\mathrm{cov}(\Theta, X)$$

- to min $\mathrm{var}(\Theta - aX)$, take deriv, set $= 0$

$$a = \frac{\mathrm{cov}(\Theta, X)}{\sigma_X^2} = \frac{\rho\sigma_\Theta\sigma_X}{\sigma_X^2} = \rho\frac{\sigma_\Theta}{\sigma_X}$$

- where $\rho = \dfrac{\mathrm{cov}(\Theta, X)}{\sigma_\Theta\sigma_X}$ is Correlation Coefficient

Resulting MS estimation error of resulting linear estimator $\hat{\Theta}$ given by

$$\mathrm{Var}(\Theta - \hat{\Theta}) = \sigma_\Theta^2 + a^2\sigma_X^2 - 2a\,\mathrm{cov}(\Theta, X)$$

$$= \sigma_\Theta^2 + \rho^2\frac{\sigma_\Theta^2}{\sigma_X^2}\sigma_X^2 - 2\rho\frac{\sigma_\Theta}{\sigma_X}\rho\sigma_\Theta\sigma_X$$

$= (1-\rho^2)\sigma_\Theta^2$

Summary LLMS

$$\hat{\Theta} = E[\Theta] + \frac{cov(\Theta, X)}{var(X)}(X - E[X])$$

$$= E[\Theta] + \rho\frac{\sigma_\Theta}{\sigma_X}(X - E[X])$$

$\rho = \frac{cov(\Theta, X)}{\sigma_\Theta \sigma_X}$ correlation coefficent

Resulting MS estimation error $= (1-\rho^2)\sigma_\Theta^2$

- easy to calculate since only mean, var, cov (Θ, X)
- more intuitive

Multiple Observations + Multiple Parameters

analgeous to LMS

(skipping section)

Linear Estimation + Normal Models

- different and inferior from LMS
- but if it happens to be linear - coincides w/ LLMS (really!?! -sarcasm)

When estimation of normal RV θ based on observations $X_i = \theta + W_i$

... ind, 0 mean, noise terms

Same as if original RVs were normal

Choice of Variables in Linear Estimation

- if transform $Y_i = h(X_i)$ $i = 1, \ldots n$
- where Y is $1 \to 1$ then has same info
- so LMS estimator is same
- what you draw may not be good linear estimate
- so subtly transform (confused)

end of chap

12/1

LECTURE 23

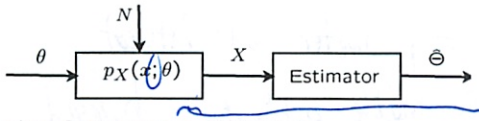
Replacement prof

- **Readings:** Section 9.1
  (not responsible for t-based confidence intervals, in pp. 471-473)

- **Outline**
  - Classical statistics
  - Maximum likelihood (ML) estimation
  - Estimating a sample mean
  - Confidence intervals (CIs)
  - CIs using an estimated variance

Starting chap 9
- big breakpoint
- 8+9 both about inference
- Changing modeling → unknown param vs known constants
- classical stats

Classical statistics



modeling framework
Θ = param of interest
still have probilistic model
Use a semicolon

- also for vectors X and θ:
  $p_{X_1,\ldots,X_n}(x_1,\ldots,x_n;\theta_1,\ldots,\theta_m)$

- These are NOT conditional probabilities; θ is NOT random
  - mathematically: many models, one for each possible value of θ

from every θ, have a model for X
↳ observation model
$\hat{\Theta}$ estimate still a RV

- **Problem types:**
  - Hypothesis testing:
    $H_0: \theta = 1/2$ versus $H_1: \theta = 3/4$
  - Composite hypotheses:
    $H_0: \theta = 1/2$ versus $H_1: \theta \neq 1/2$
  - Estimation: design an **estimator** $\hat{\Theta}$, to keep estimation **error** $\hat{\Theta} - \theta$ small

Can address same issues as w/ Bayes stats

Maximum Likelihood Estimation

Lot of criteria to try to estimate for

- Model, with unknown parameter(s): $X \sim p_X(x;\theta)$
- Pick θ that "makes data most likely"

$$\hat{\theta}_{\text{ML}} = \arg\max_\theta p_X(x;\theta)$$

value for θ that makes x we observed most likly

- Compare to Bayesian MAP estimation:

$$\hat{\theta}_{\text{MAP}} = \arg\max_\theta \frac{p_{X|\Theta}(x|\theta)p_\Theta(\theta)}{p_X(x)} = \arg\max_\theta p_{\Theta|X}(\theta|x)$$

maximize posterior definition

- **Example:** $X_1, \ldots, X_n$: i.i.d., exponential(θ)

"drawn from" "don't know"

$$\max_\theta \prod_{i=1}^n \theta e^{-\theta x_i}$$

joint probability

$$\max_\theta \left(n \log\theta - \theta \sum_{i=1}^n x_i \right)$$

want to maximize over the choices for θ

$$f_{X_i}(x;\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Or instead max log of the function

$$\hat{\theta}_{\text{ML}} = n/(x_1 + \ldots + x_n)$$

use cap. letters

$$\hat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n}$$ RV

reciprocal of n RVs

take deriv, set = 0, max is

$$\frac{d}{d\theta} \to \frac{n}{\theta} - \sum_{i=1}^{\hat{}} X_i$$

Very important estimator
- Use freq
- but only 1 possible estimator

Desirable properties of estimators

(should hold FOR ALL θ !!!)

- **Unbiased:** $\mathrm{E}[\hat{\Theta}_n] = \theta$ expected value of estimator should = parameter of interest
  - exponential example, with $n = 1$: $\mathrm{E}[1/X_1] = \infty \neq \theta$ (biased) — other sheet
  - Many reasonable estimators biased
- **Consistent:** $\hat{\Theta}_n \to \theta$ (in probability) diff way to look at estimator
  - exponential example: $(X_1 + \cdots + X_n)/n \to \mathrm{E}[X] = 1/\theta$ by weak law large # sample mean converges to the mean
  - can conclude that: $\hat{\Theta}_n = n/(X_1 + \cdots + X_n) \to 1/\mathrm{E}[X] = \theta$
- **"Small" mean squared error (MSE)**

$$\begin{aligned} \mathrm{E}[(\hat{\Theta} - \theta)^2] &= \mathrm{var}(\hat{\Theta} - \theta) + (\mathrm{E}[\hat{\Theta} - \theta])^2 \\ &= \mathrm{var}(\hat{\Theta}) + (\text{bias})^2 \end{aligned}$$

will be some function of θ - non random

General problem

Estimate a mean

- $X_1, \ldots, X_n$: i.i.d., mean θ, variance σ^2

$X_i = \theta + W_i$ ↑ seq RV ↑ fixed param of interest

W_i: i.i.d., mean, 0, variance σ^2

$\hat{\Theta}_n$ = sample mean $= M_n = \dfrac{X_1 + \cdots + X_n}{n}$

Properties:

- $\mathbf{E}[\hat{\Theta}_n] = \theta$ (unbiased) ← elementary properties of expectation
- WLLN: $\hat{\Theta}_n \to \theta$ (consistency)
- MSE: σ^2/n
- Sample mean often turns out to also be the ML estimate.
  E.g., if $X_i \sim N(\theta, \sigma^2)$, i.i.d.

$$\mathbf{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] = \frac{1}{n}\left[\mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n]\right]$$
$$= \frac{1}{n} \cdot n\theta = \theta$$

↑ can compute $\mathrm{MSE} = \mathbf{E}[(\hat{\theta}_n - \theta)^2] = \mathbf{E}\left[\left(\frac{X_1 + \cdots + X_n}{n} - \theta\right)^2\right] = \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right)$

$$= \frac{1}{n^2}\left(\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)\right)$$
$$= \frac{\sigma^2}{n}$$

↑ many cases max likelyhood estimates

Confidence intervals (CIs)

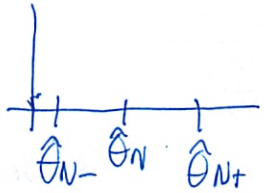
How do we know if estimate is a good one?
Very different from Bayes
- Seperate sheet

- An estimate $\hat{\Theta}_n$ may not be informative enough
- An $1 - \alpha$ **confidence interval** is a (random) interval $[\hat{\Theta}_n^-, \hat{\Theta}_n^+]$,

  s.t. $\mathbf{P}(\hat{\Theta}_n^- \le \theta \le \hat{\Theta}_n^+) \ge 1 - \alpha, \quad \forall\, \theta$

  - often $\alpha = 0.05$, or 0.25, or 0.01
  - interpretation is subtle
- CI in estimation of the mean $\hat{\Theta}_n = (X_1 + \cdots + X_n)/n$
  - normal tables: $\Phi(1.96) = 1 - 0.05/2$

Want interval



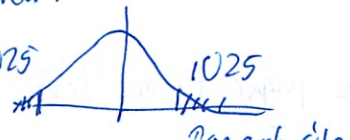
True θ not random

$$\mathbf{P}\left(\frac{|\hat{\Theta}_n - \theta|}{\sigma/\sqrt{n}} \le 1.96\right) \approx 0.95 \quad \text{(CLT)}$$

↑ normal tables

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{1.96\,\sigma}{\sqrt{n}} \le \theta \le \hat{\Theta}_n + \frac{1.96\,\sigma}{\sqrt{n}}\right) \approx 0.95$$

↑ So prob of this is $\approx .95$



↑ on each side

More generally: let z be s.t. $\Phi(z) = 1 - \alpha/2$

$$\mathbf{P}\left(\hat{\Theta}_n - \frac{z\sigma}{\sqrt{n}} \le \theta \le \hat{\Theta}_n + \frac{z\sigma}{\sqrt{n}}\right) \approx 1 - \alpha$$

The case of unknown σ

Pollster problem again

- Option 1: use upper bound on σ
  - if X_i Bernoulli: $\sigma \le 1/2$
- Option 2: use ad hoc estimate of σ
  - if X_i Bernoulli(θ): $\hat{\sigma} = \sqrt{\hat{\Theta}(1-\hat{\Theta})}$
- Option 3: Use generic estimate of the variance
  - Start from $\sigma^2 = \mathrm{E}[(Y-\theta)^2]$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\theta)^2 \to \sigma^2$$

(but do not know θ)

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\hat{\Theta}_n)^2 \to \sigma^2$$

(unbiased: $\mathrm{E}[\hat{S}_n^2] = \sigma^2$)

bounds on var - usually ty for Bernoulli

Estimate fraction that hold particular opinion
People who vote yes on a particular example
Markov + Chebyshev for crude bound
CLT for more exact

Size of CI also depends on σ
Need to form estimate for var() to use in estimating CI
- using option 3 → form estimated var

sample mean
Convergence in prob by WLLN
but need θ

Slight adjustment: multiply natural estimate
$S_n^2 = \frac{n}{n-1} \cdot$ (more obvious estimate)

An example of an exact CI

Another example where have definite CI

- X: exponential with parameter θ
- Analyze $[a/X, b/X]$ as a confidence interval for θ: ← reasonable estimate we considered previously

$$\begin{aligned}\mathrm{P}\left(\frac{a}{X} \le \theta \le \frac{b}{X}\right) &= \mathrm{P}\left(\frac{a}{\theta} \le X \le \frac{b}{\theta}\right) \\ &= \int_{a/\theta}^{b/\theta} \theta e^{-\theta x}\,dx \\ &= -e^{-\theta x}\Big|_{x=a/\theta}^{x=b/\theta} = e^{-a} - e^{-b}\end{aligned}$$

← elementary computation

- we know how to compute $P(X \in [\alpha, \beta])$
- rearrange so prob is in that form

$= \int_{a/\theta}^{b/\theta} f_X(x;\theta)\,dx$

No dependence on θ, so have a confidence interval

- Example: $\left[\frac{1}{4X}, \frac{4}{X}\right]$ is a 0.76 confidence interval ("76% CI")

Confidence interval
no dependence on θ

if tried $[a\sqrt{X}, b\sqrt{X}]$ - would not have gotten prob ind. of θ

* Much chose proper form "trick" to find CI not dependent on θ - could use instead CLT and bound on var

Exponential example

$$\hat{\theta}_n = \frac{n}{X_1 + \cdots + X_n}$$

$$\hat{\theta}_1 = \frac{1}{X_1}$$

$$E[\hat{\theta}_1] = E\left[\frac{1}{X_1}\right]$$

$$= \int_0^\infty \frac{1}{x} \theta e^{-\theta x} dx$$

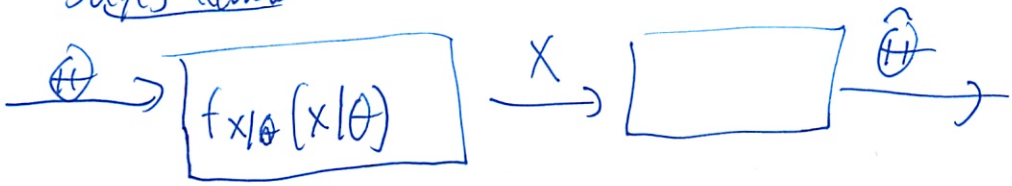
~~$= \theta$~~ No

The $\int$ diverges for all θ values of θ

$= \infty$

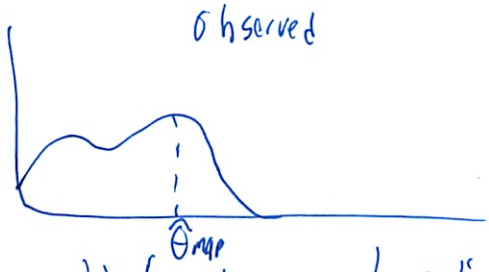
So this one is biased

(2) Bayes Review



$\uparrow f_{\Theta|X}(\theta|x)$

? little x fixed since observed



Boil down dist to 1 estimate, like the $\hat{\theta}_{MAP}$ estimate

Can ask qu abat quality of estimate

$E[\downarrow(\Theta, \hat{\Theta}_{MAP})]$ Can compute error

---

But classical setting different θ is <u>not</u> random

Can give a confidence interval

Michael Plasmeier

12/2

4/10

1.5
2.0
0.5

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Problem Set 10
Due December 2, 2010 Thur

1. **A financial parable.** An investment bank is managing \$1 billion, which it invests in various financial instruments ("assets") related to the housing market (e.g., the infamous "mortgage backed securities"). Because the bank is investing with borrowed money, its actual assets are only \$50 million (5%). Accordingly, if the bank loses more than 5%, it becomes insolvent. (Which means that it will have to be bailed out, and the bankers may need to forgo any huge bonuses for a few months.)

   (a) The bank considers investing in a single asset, whose gain (over a 1-year period, and measured in percentage points) is modeled as a normal random variable R, with mean 7 and standard deviation 10. (That is, the asset is expected to yield a 7% profit.) What is the probability that the bank will become insolvent? Would you accept this level of risk?

   (b) In order to safeguard its position, the bank decides to diversify its investments. It considers investing \$50 million in each of twenty different assets, with the ith one having a gain R_i, which is again normal with mean 7 and standard deviation 10; the bank's gain will be $(R_1 + \cdots + R_{20})/20$. These twenty assets are chosen to reflect the housing sectors at different states or even countries, and the bank's rocket scientists choose to model the R_i as independent random variables. According to this model, what is the probability that the bank becomes insolvent?

   (c) Based on the calculations in part (b), the bank goes ahead with the diversified investment strategy. It turns out that a global economic phenomenon can affect the housing sectors in different states and countries simultaneously, and therefore the gains R_i are in fact positively correlated. Suppose that for every i and j where $i \neq j$, the correlation coefficient $\rho(R_i, R_j)$ is equal to $1/2$. What is the probability that the bank becomes insolvent? You can assume that $(R_1 + \cdots + R_{20})/20$ is normal.

2. The adult population of Nowhereville consists of 300 males and 196 females. Each male (respectively, female) has a probability of 0.4 (respectively, 0.5) of casting a vote in the local elections, independently of everyone else. Find a good numerical approximation for the probability that more males than females cast a vote.

3. Let S_n be the number of successes in n independent Bernoulli trials, where the probability of success in each trial is $p = \frac{1}{2}$. Provide a numerical value for the limit as n tends to infinity for each of the following three expressions:

   (a) $\mathbf{P}(\frac{n}{2} - 10 \leq S_n \leq \frac{n}{2} + 10)$

   (b) $\mathbf{P}(\frac{n}{2} - \frac{n}{10} \leq S_n \leq \frac{n}{2} + \frac{n}{10})$

   (c) $\mathbf{P}(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq S_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2})$

4. Alice has two coins. The probability of heads for the first coin is $1/3$; the probability of heads for the second coin is $2/3$. Other than this difference in their bias, the coins are indistinguishable through any measurement known to man. Alice chooses one of the coins randomly and sends it to Bob. Let p be the probability that Alice chose the first coin. Bob tries to guess which of the two coins he received by flipping it 3 times in a row and observing the outcome. Assume that all coin flips are independent. Let Y be the number of heads Bob observed.

(a) Given that Bob observed k heads, what is the probability that he received the first coin?

(b) Find values of k for which the probability that Alice sent the first coin increases after Bob observes k heads out of 3 tosses. In other words, for what values of k is the probability that Alice sent the first coin given that Bob observed k heads greater than p? If we increase p, how does your answer change (goes up, goes down, or stays unchanged)?

(c) Help Bob develop the rule for deciding which coin he received based on the number of heads k he observed in 3 tosses if his goal is to minimize the probability of error.

(d) For this part, assume $p = 2/3$.

i. Find the probability that Bob will guess the coin correctly using the rule above.

ii. How does this compare to the probability of guessing correctly if Bob tried to guess which coin he received before flipping it?

(e) If we increase p, how does that affect the decision rule?

(f) Find the values of p for which Bob will never guess he received the first coin, regardless of the outcome of the tosses.

(g) Find the values of p for which Bob will always guess he received the first coin, regardless of the outcome of the tosses.

5. Consider a Bernoulli process $X_1, X_2, X_3, \ldots$ with unknown probability of success q. As usual, define the kth inter-arrival time T_k as

$$T_1 = Y_1, \qquad T_k = Y_k - Y_{k-1}, \quad k = 2, 3, \ldots$$

where Y_k is the time of the kth success. This problem explores estimation of q from observed inter-arrival times $\{t_1, t_2, t_3, \ldots\}$.

You may find the following integral useful: For any non-negative integers k and m,

$$\int_0^1 q^k (1-q)^m dq = \frac{k!\, m!}{(k+m+1)!}$$

Assume q is sampled from the random variable Q which is uniformly distributed over $[0, 1]$.

(a) Compute the PMF of T_1, $p_{T_1}(t_1)$

(b) Compute the least squares estimate (LSE) of Q from the first recording $T_1 = t_1$.

(c) Compute the maximum a posteriori (MAP) estimate of Q given the k recordings, $T_1 = t_1, \ldots, T_k = t_k$.

For this part only assume q is sampled from the random variable Q which is now uniformly distributed over $[0.5, 1]$

(d) Find the linear least squares estimate (LLSE) of the second inter-arrival time (T_2), from the observed first arrival time ($T_1 = t_1$).

6. The joint PDF of X and Y is defined as follows:

$$f_{X,Y}(x,y) = \begin{cases} cxy & \text{if } 0 < x \leq 1,\ 0 < y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

(a) Find the normalization constant c.

(b) Compute the conditional expectation estimator of X based on the observed value $Y = y$.

(c) Is this estimate different from what you would have guessed before you saw the value $Y = y$? Explain.

(d) Repeat (b) and (c) for the MAP estimator.

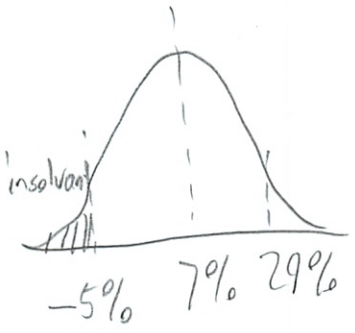
OH P - Set nothing new
- just estimators
- Bayes rule

Michael Plasmeier P-Set 10 12/2

1. A financial parable — Investment bank

- managing $1 billion in various housing assets
- only $50 million / 5% actual assets
- so if it loses more than 5% it may be insolvent

a) Bank considers investing in a single asset
- gain over 1 year in % points is normal RV R w/ $\mu = 7$ $\sigma = 10$
  - Expected to yield a 7% profit
- What is prob that bank will become insolvent?



- ? Simple limit theorem / lecture 19 problem
- Well depends which method one uses
  - Chebshev
  - Markov

- No! One RV, so can just use CDF

$$P(R \leq -5\%) = \Phi\left(\frac{R - E[X]}{\sqrt{Var(X)}}\right)$$

$Var = \sigma^2$

$\sqrt{Var} = \sigma$

(2)

$= \Phi\left(\frac{R-7}{10}\right)$

$= \Phi\left(\frac{-5-7}{10}\right)$

$= \Phi(-1.2)$

$= 1 - \Phi(1.2)$

$= 1 - .8849$

$= .1151$

11% risk of failure

- a bit more than 1 st dev which is ~15% - so makes sense

Would I bet the firm on this?

No - I am a more conservative investor.

b) In order to safeguad its position, bank diversifies $50 million in 20 different assets

R_i = ith one's gain

↳ each one mean 7 st dev 10

Bank's total gain rate = $\frac{R_1 + \cdots + R_{20}}{20} = \rho$

bank's "rocket scientists" model each R_i as ind

Now what is prob bank is insolvent.

(I like this problem)

(3)

Here limit theory of sums

figuring how to do grader skip pg

- Chebchiev ?
- Markov ?
- Something more exact?
- Central Limit Theorm
  - most exact
  - is it perfectly exact or just more exact?
- Is like Recitation 21 #1
  - CLT is actual estimate, not exact though
  - Chebishev + Markov are upper bounds
- How do actual again?

WP: Convolution of their densities

↳ aka cross correlation

the shifting function

↳ common area

TB: Sum independent RVs

(4)

$$P(P \leq -5) = P\left(\frac{P - E[R]}{\sqrt{Var(R)}} \leq \frac{-5 - E[R]}{\sqrt{Var(R)}}\right)$$

$$= P\left(\frac{P - E[R]}{\sqrt{Var(R)}} \leq \frac{-5 - 7}{10}\right)$$

$$= \Phi(-1.2)$$

$$= 1 - \Phi(1.2)$$

$$= 1 - .8849$$

$$= .1151$$

So same as before

Was not expecting, but makes sense since noise on both sides of mean

This is for ∞ n I believe

well $\lim_{n \to \infty}$

(5)

$S_n = P$ here

$$Z_n = \frac{S_n - E[S_n]}{\sigma_{S_n}}$$

$$= \frac{S_n - nE[X]}{\sqrt{n}\,\sigma}$$

ebb! when sum of RVs

$$P(P \leq -5) = P\left(\frac{P - nE[R_i]}{\sqrt{n}\sqrt{Var(R_i)}} \leq \frac{-5 - nE[R_i]}{\sqrt{n}\,\sigma_{R_i}}\right)$$

$$= P\left(\frac{P - nE[R_i]}{\sqrt{Var(R_i)}} \leq \frac{-5 - 20 \cdot 7}{\sqrt{20} \cdot 10}\right)$$

$$= \Phi(-3.24)$$

Well wait $M_n = R$

and $S_n = nM_n$

$$= 1 - \Phi(3.24)$$

$$= 1 - .9994$$

$$= .0006$$

Bank will almost certainly not go bankrupt

Much more believable outcome, though I would not think that small

(6)

Remember Chebyshev better than Markov!

c) Based on b) bank will diversify. But turns out RVs are not independent → positively correlate.

Suppose that for every i and j where $i \neq j$ the correlation coefficient $\rho(R_i, R_j) = \frac{1}{2}$

What is prob bank will go insolvent?

? Did we ever study this?

- CLT only when ind
- Markov + Cheb, dep or ind
- but Cheb allows −, so can use
  - p220 chap 4
  - must be careful to compute sum
  - sum of var + sum of cov
    - both ways must count
  - can derive, not in book

$$\text{var}(X_1 + \cdots + X_n) = E[(X_1 + \cdots + X_n)^2] - E[X_1 + \cdots + X_n]^2$$

Should simplify to expression w/ vars + cov

then from correlation coefficient, to cov

- last sentance gives us a critical assumption
  - that is normal (else no CLT) → just normal CDF

(7)

$\rho(X,Y) = \frac{1}{2}$

$\Rightarrow$ by definition

$\frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$ can back at cov

↑ mnemonic $\frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

$Var(X) = \sigma_X^2$

↑ st dev

$\sigma_{XY} \leftarrow Cov(X,Y)$

Since $Var(X) = Cov(X,X)$

$E\left[\left(\sum_{i=1}^{n} X_i\right)^2 + \right.$

↑ like terms

... all shown on p 220

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{(i,j \mid i \neq j)} Cov(X_i, X_j)$$

↑ double summation

$$\sum_{i=1}^{20} \sum_{\substack{j=1 \\ i \neq j}}^{20}$$

So $Cov(X_1, X_2) + Cov(X_1, X_3) + \ldots + Cov(X_2, X_1) + Cov(X_2, X_3)$

↑ no $Cov(X_1, X_1)$ ↑ double count

(8)

Back out correlation coefficient

$$\frac{1}{2} = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$$

$$Var(R) = \sigma^2 = 10^2$$

$$\sqrt{Var} = \sigma = 10$$

$$\frac{1}{2} = \frac{Cov(X,Y)}{\sqrt{100 \cdot 100}}$$

$$50 = Cov(X,Y)$$

$$Var\left(\sum_{i=1}^{20} X_i\right) = \sum_{i=1}^{20} 100 + \sum_{(i,j \mid i \neq j)}^{20} 50$$

$$= 20 \cdot 100 + 19 \cdot 19 \cdot 50$$

$$= 20{,}050$$

? Var for each

$E[\,]$ unchanged?

(9)

$$P(P \le -5) = P\left(\frac{P - nE[R_i]}{\sqrt{n}\sqrt{Var(R_i)}} \le \frac{-5 - 20 \cdot 7}{\sqrt{20}\sqrt{20050}}\right)$$

$$= \Phi(-.2289)$$

ha - actually worked out fairly well

$$= 1 - \Phi(.2289)$$

$$= 1 - .5871$$

$$= .4129$$

Makes sense

- much higher
- but is it too high?

thankfully can't do this comply on final

(10)

2. Adult population of Nowhere ville

= 300 males , 4 prob to vote

= 196 females , 5 prob to vote

Find a good # approx for prob more males vote than females

M_i = prob ith man votes $\begin{cases} 1 \text{ yes} & \begin{cases} .4 \text{ male} \\ .6 \text{ "} \end{cases} \\ 0 \text{ no} & \begin{cases} .5 \text{ female} \\ .5 \text{ "} \end{cases} \end{cases}$

F_i = " " female "

$M = M_1 + M_2 + \dots + M_i$ ← sum of ind Bernoullis = Binomial

$F = F_1 + F_2 + \dots + F_i$

$P(M > F)$

- So comparing 2 RVs

Can do p certain and will vote, say 21

exact $P(S_n \leq 21) = \sum_{k=0}^{21} \binom{300}{k} (.4)^{300}$ ← because binomial

CLT $\approx P\left(\frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} \leq \frac{21 - 300 \cdot .4}{\sqrt{300 \cdot .4(1-.4)}} \right)$

$E[S_n] = np$, $Var(S_n) = np(1-p)$) Binomial

$\approx \Phi(-1.35)$

but then usually add $\frac{1}{2}$

(11)

But want $P(M > F) = 1 - P(M \leq F)$

$$1 - P\left(\frac{M - E[M]}{\text{Var}(M)} \leq \frac{F - E[F]}{\text{Var}(F)}\right)$$

← note this is S_n - so its not $nE[X]$

$E[M] = np = 300 \cdot .4 = 120$

$\text{Var}(M) = np(1-p) = 300 \cdot .4(1-.4) = 72$

$E[F] = np = 196 \cdot .5 = 98$

$\text{Var}(F) = np(1-p) = 196 \cdot .5 \cdot (1-.5) = 49$

$$1 - P\left(\frac{M - 120}{72} \leq \frac{F - 98}{49}\right)$$

? so now need dist of this

well want a single #

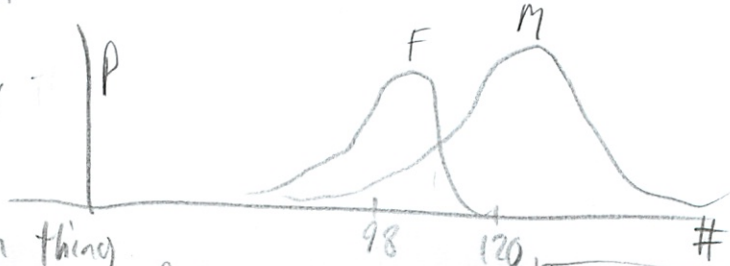
but how translate to std normal table

? did we ever do a problem like this?

Or do something else

+0.5

- both are normal distributions
- want overlap
- No want prob of overlap
- well are prob dist
- but what graphically is the ans?



- ? some weird convolution thing
- or add them into some new RV

Tile slowly piecing together

(12)

Sum of ind Bernolis - almost normal

~~Could figure Mean, Var M, F~~

So could do joint - had

So do $V = M - F$

and find distribution $P(V \geq 0)$

+0.5

duh - should have thought of this - was leaning towards

- sum of ind normal is normal
  - or any linear comb

CLT - sum of any iid, standardized are approx normal

$E[V]$ = expectation of sum of RV

linearity of expectation

$= E[M] - E[F]$

$= 120 - 98$

$= 22$ +0.5

$Var(V)$ = oh right was looking for this on Nov 20 P-set

$= Var(X) + Var(Y) + 2Cov(X, Y)$

find so 0

$= 72 + 49$

$= 121$

$$P(V \geq 0) = P\left(\frac{V - E[V]}{Var(V)} \geq \frac{0 - 22}{121}\right)$$ (-0.5)

(13)

$= \Phi(-.18)$

$= 1 - \Phi(.18)$

$= 1 - .5714$

$= .4286$

✓ Makes sense

1.5/2

Challenging P-set

3. Let S_n be # of sucesses in n ind. Bernalli trials where $p(\text{sucess}) = \frac{1}{2}$

Provide a numerical valve as $n \to \infty$

This is <u>DeMoivre - Laplace Approx to Binomial</u>

- we did not go over much

$$\mu = E[X_i] = p = \frac{1}{2} \qquad \sigma = \sqrt{\text{var}(X_i)} = \sqrt{p(1-p)} = \sqrt{\frac{1}{2}\left(1-\frac{1}{2}\right)} = \frac{1}{2}$$

We will use approx from central limit theorm to provide approx for the prob $\{k \le S_n \le l\}$

Convert into a standardized RV using the equivilance

$$\frac{k-np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{l-np}{\sqrt{np(1-p)}}$$

$$P(k \le S_n \le l) = P(\ \nearrow\ )$$

$$= \Phi\left(\frac{l-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$ then add $\frac{1}{2}$ for better estimation

CDF of one - the other

(15)

a) $P\left(\frac{n}{2} - 10 \le S_n \le \frac{n}{2} + 10\right)$

as $n \to \infty$ this gets larger

the n in there throws a wrench in my plans

$$\Phi\left(\frac{\frac{n}{2} + 10 + \frac{1}{2} - n\frac{1}{2}}{\sqrt{n\,\frac{1}{2}\left(\frac{1}{2}\right)}}\right) - \Phi\left(\frac{\frac{n}{2} - 10 - \frac{1}{2} - n\frac{1}{2}}{\sqrt{n\,\frac{1}{2}\left(\frac{1}{2}\right)}}\right)$$

$$\Phi\left(\frac{5}{\sqrt{\frac{n}{4}}}\right) - \Phi\left(\frac{-5}{\sqrt{\frac{n}{4}}}\right) \qquad \sqrt{\frac{n}{4}} = \frac{\sqrt{n}}{2}$$

As $n \to \infty$, denominator grows to ∞

So $\Phi(\,) \to 0$, so value $\to$ (15)

(−1)

(16)

b) $P\left(\frac{n}{2} - \frac{n}{10} \leq S_n \leq \frac{n}{2} + \frac{n}{10}\right)$

$$\Phi\left(\frac{\frac{n}{2} + \frac{n}{10} + \frac{1}{2} - n\frac{1}{2}}{\sqrt{n\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}}\right) - \Phi\left(\frac{\frac{n}{2} - \frac{n}{10} - \frac{1}{2} - n\frac{1}{2}}{\sqrt{n\frac{1}{2}\left(\frac{1}{2}\right)}}\right)$$

$$\Phi\left(\frac{\frac{n}{10} + \frac{1}{2}}{\sqrt{\frac{n}{4}}}\right) - \Phi\left(\frac{-\frac{n}{10} - \frac{1}{2}}{\sqrt{n\frac{1}{2}\left(\frac{1}{2}\right)}}\right)$$

So again denom $\to \infty$) as $n \to \infty$

But now numerator also goes $\to \infty$

So $\Phi\left(\frac{\infty}{\infty}\right)$ which is indeterminate

But put whole thing into wolfram alpha $\Phi(\,) \to \infty$

Which leads value $\to 1$

Other side $\Phi(\,) \to -\infty$

So this is $1 - \Phi(\infty)$ which is $1 - 1 = 0$

So $1 - 0 = \boxed{1}$

+1

These are cool

(17)

c) $P\left(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq S_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2}\right)$

$$\Phi\left(\frac{\frac{n}{2} + \frac{\sqrt{n}}{2} + \frac{1}{2} - n\frac{1}{2}}{\sqrt{n\frac{1}{2}\left(\frac{1}{2}\right)}}\right) - \Phi\left(\frac{\frac{n}{2} - \frac{\sqrt{n}}{2} - \frac{1}{2} - n\frac{1}{2}}{\sqrt{n\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}}\right)$$

$$\Phi\left(\frac{\frac{\sqrt{n}}{2} + \frac{1}{2}}{\sqrt{\frac{n}{4}}}\right) - \Phi\left(\frac{\frac{-\sqrt{n}}{2} - \frac{1}{2}}{\sqrt{\frac{n}{4}}}\right)$$

Denom still $\to \infty$

Numerator $\to \infty$

Which is indeterminate

But whole thing together $\to 1$ (wolfram alpha)

So $\Phi(1) \to .8413$

Other one together $\to -1$

So $1 - \Phi(1) = 1 - .8413$

So $.8413 - (1 - .8413)$

.16826

+1

2/3

(18)

4. Alice has 2 coins

$P(\text{heads 1st coin}^{A}) = \frac{1}{3}$

$P(\text{heads 2nd coin}^{B}) = \frac{2}{3}$

Alice chooses coin randomly + sends to Bob.

P = prob Alice picked first coin
$(\frac{1}{2}$, right)

Bob tries to guess by flipping his coin 3x

X = # of heads he sees.

Simple Basian problem

Use CLT
then lim as $n \to \infty$

θ = which coin Chosen

P_θ $\begin{cases} \frac{1}{2} & \text{coin A} \\ \frac{1}{2} & \text{coin B} \end{cases}$ - well $\begin{cases} P & \text{coin A} \\ 1-P & \text{coin B} \end{cases}$

But it said coins indinguished, so we can assume $p = \frac{1}{2}$

$P_{X|\theta} = \begin{cases} \text{for } \theta = A \\ \frac{1}{3} \text{ heads} \\ \frac{2}{3} \text{ tails} \end{cases} = \begin{cases} \text{for } \theta = B \\ \frac{2}{3} \text{ heads} \\ \frac{1}{3} \text{ tails} \end{cases}$

Just an application of Bayes Rule
- define RV or events
- (Alice gives bob coin B)

OH

In general $P(A_i | Y=k) =$ ← # heads you got

↑ got 1st coin

$$= \frac{P(A_i)\, P(Y=k|A_i)}{\sum_{i=1}^{n} P(A_i)\, P(Y=k \mid A_i)} \leftarrow P(Y=k)$$

total prob theorem

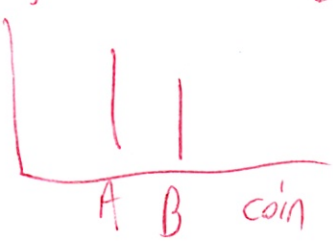
$\Theta = \begin{cases} 1 & \text{1st coin} \\ 2 & \text{2nd coin} \end{cases}$ ← just syntax

Not implied that $p = \frac{1}{2}$

Which estimator MAP, LMS, LLMS

↑ not good for our purposes

$P_{\Theta|Y}(\theta|k)$ ← just have 2 #



(oh so I was on right track!)

So conditional prob of error = prob of other

So just do $\hat{\theta}_{MAP} = \arg\max P_{\Theta|X}(\theta|x)$

Depends on p in C

↑ $\frac{P_\Theta(\theta)\, P_{X|\Theta}(x|\theta)}{\sum P_\Theta(\theta)\, P_{X|\Theta}(x|\theta)}$ ← don't have to worry about denom

(20)

$P_\theta(1)\, P_{X|\theta}(x|1) \geq P_\theta(2)\, P_{X|\theta}(x|2)$

↑ pick coin A

- in terms of k, p
- now it give p - what it will be
- but can do which is bigger w/ fixed p

b) ~~right~~ side in terms of k
both sides in terms of p
for which values of k is this true

(21)

? into chap 8 now

a) Given that he saw k heads, prob that recieved 1st coin

Y is binomial if $\begin{cases} \theta = A & p = \frac{1}{3} \\ \theta = B & p = \frac{2}{3} \end{cases}$ $n = 3$

So must chose b/w these two models

Now what is the simple math for this?

-remember 6.01

$$f_{\theta|x}(\theta|x) = \frac{f_\theta(\theta) f_{x|\theta}(x|\theta)}{\sum f_\theta(\theta) f_{x|\theta}(x|\theta)}$$

So first if $\theta = A$

$$\frac{1}{2} \cdot \sum_{k=0}^{3} \binom{3}{k} \frac{1}{3}^k \left(\frac{2}{3}\right)^{3-k}$$

made bad assumption, $p = \frac{1}{2}$

$$\frac{1}{2} \cdot \left(3 \cdot \frac{1}{3}\left(\frac{2}{3}\right)^2 + \frac{3!}{2!1!} \cdot \left(\frac{1}{3}\right)^2 \left(\frac{2}{3}\right) + 1 \cdot \left(\frac{2}{3}\right)^3\right)$$

$$\frac{1}{2}\left(\frac{4}{9} + \frac{2}{9} + \frac{8}{27}\right)$$

$$\frac{1}{2}\left(\frac{12}{27} + \frac{6}{27} + \frac{8}{27}\right)$$

$$\frac{1}{2}\left(\frac{26}{27}\right)$$

$\theta = B$

$$\frac{\frac{1}{2} \cdot \sum_{k=0}^{3} \binom{3}{k} \frac{2}{3}^k \left(\frac{1}{3}\right)^{3-k}}{\frac{1}{2}\left(3 \cdot \frac{2}{3}\left(\frac{1}{3}\right)^2 + 3 \cdot \left(\frac{2}{3}\right)^2\left(\frac{1}{3}\right) + 1\left(\frac{1}{3}\right)^3\right)}$$

$$\frac{1}{2}\left(\frac{1}{27}\right)$$

$$\frac{1}{2}\left(\frac{1}{27}\right)$$

$$\frac{1}{2}\left(\frac{6}{27} + \frac{6}{27} + \frac{1}{27}\right)$$

$$\frac{1}{2}\left(\frac{13}{27}\right)$$

$$\frac{1}{2}\left(\frac{13}{27}\right)$$

$$\frac{\frac{1}{2}\left(\frac{1}{27}\right)}{\frac{1}{2}\left(\frac{13}{27}\right)} = \frac{1}{13}$$

(28)

$$\frac{1}{2}\left(\frac{2}{9} + \frac{4}{9} + \frac{1}{27}\right)$$

$$,3518$$

Sum $,481 + ,3518 = ,83 = \frac{5}{6}$

So $\Theta = \begin{cases} A & ,57 \\ B & ,42 \end{cases}$

There - I think that was the manually way

But what math shortcuts are there?

But wait the answer should be based on k heads

So what did I do?

Possibilities for θ, for <u>any/all</u> values of k

So as function of k

$$\frac{\frac{1}{2} \cdot \binom{3}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{3-k}}{\binom{3}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{3-k} + \binom{3}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{3-k}}$$

And this is only for 1st one

After OH

So ,57 was right for map

$\hat{\Theta}_{MAP} = A$

(23)

b) Find values for k for which the p that Alice sent first coin increases after Bob observes k heads out of 3 tosses,

Aka for what values of k is the p(Alice sent 1st coin | B observed k heads) $> p$

If $\neq p$, how does answer change.

So if $k = 0$, then likely to have 1st coin "A"
$k = 3$, " " " " 2nd coin "B"

But specifically what is prob

So for what values of k is

or not $\frac{1}{2}$ $\rightarrow$ $\dfrac{p\binom{3}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{3-k}}{\binom{3}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{3-k} + \binom{3}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{3-k}} > p$

So $k = 0$ 2, 8
1 2, 6
2 2, 3
3 2, 1

w/ $p = 1$

Since p is on the same side, changing it won't affect anything.

(24)

No - that does not make sense.

And tested it when $p = .25$ $p = .5$

k	$p = .25$	$p = .5$
0	~.22	~.45
1	~.16	~.32
2	~.09	~.14
3	~.01	~.07

? No this is wrong - answer is always larger than p

That is just numeric value for left hand side

? Supposed to be a PMF

Perhaps should do via an estimator

After OH

For which values of k is this true

? None of them - but is not right

Recon in wolfram alpha

$$p \cdot 9 \cdot 2^{-k-1} (3-k)! \, k! > p$$

$k=0$	$27p > p$	✓	So here A is always
$k=1$	$\frac{9}{2}p > p$	✓	> than p - should
$k=2$	$\frac{9}{4}p > p$	✓	not be true
$k=3$	$\frac{27}{8}p > p$	✓	- asymptotal to 0

(25)

Try out estimating function

So we have $Y = k$ $\qquad Y = F_1 + F_2 + F_3 = S_n$

binomial, bernoulli

So we have sum $\Rightarrow$ CLT?

$$Z_n = \frac{S_n - E[S_n]}{\text{Var}(S_n)}$$

$E[S_n] = np$ ← but p varies w/ which one is chosen

binomial

$$\text{Var}(S_n) = np(1-p)$$

What is the problem even asking?

I remember how we did this in 6.04

$$Z_n = \frac{S_3 - 3p}{3p(1-p)}$$ ← but we are asked to find p

Unless its a combo of the two

$p_{\text{above}} = \begin{cases} \frac{1}{3} & \text{w/ prob } p \\ \frac{2}{3} & \text{w/ prob } (1-p) \end{cases}$ — average

$$p_{\text{above}} = \frac{1}{3}p + \frac{2}{3}(1-p)$$

(26)

$$Z_n = \frac{S_3 - 3\left(\frac{1}{3}p + \frac{2}{3}(1-p)\right)}{3\left(\frac{1}{3}p + \frac{2}{3}(1-p)\right)\left(1 - \left(\frac{1}{3}p + \frac{2}{3}(1-p)\right)\right)} =$$

Then can use LMS to find prob > something

But should this not now be fn of k

Well $S_3 = k$ ← actual value
↑ RV

$$P(S_3 \geq 0) = \Phi\left(\frac{0 - \dots}{\dots}\right)$$

$$= \Phi\left(\frac{-3}{p+1}\right)$$

depends on p ... - can't use tables

But can I find $\Phi\left(\frac{-3}{p+1}\right) > P(p)$

$$P(S_3 \geq 1) = \Phi\left(\frac{1 - \dots}{\dots}\right)$$

$$= \Phi(\text{Complicated})$$

(27)

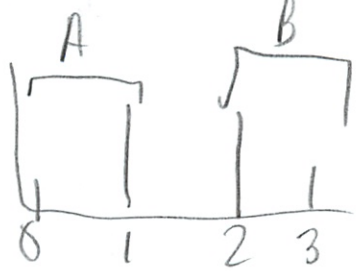
c) Help Bob develop the rule for deciding which coin he recieved based on # of heads k he sees in 3 tosses

- To minizie prob of error = MAP
  - well any estimator
- Well depends on p, but should be able to answer in terms of p
-

Perhaps just split

Since we don't know p

So



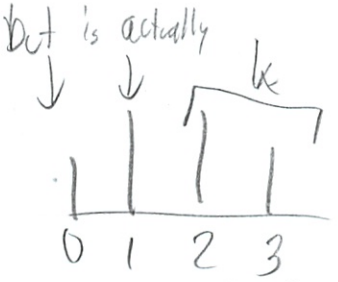
And this uses MAP → tallest

(28) I can't sort out all these variables in my head

d) Assume $p = \frac{2}{3}$

i) Find prob Bob will guess correctly

- So what guess is this again?
- After 3 observations
- Correct means after 3 data matches actual
- Wrong is error??
- So MSE??
- Error is if says A, and is B



So $p\,P(\text{B when } k=2, k=3) + (1-p)\,P(\text{B when } k=0, k=1)$

(29)

ii) How does this compare w/ prob of guessing correctly if Bob tried to guess which coin he rolled before flipping it.

Well before flipping it, he only knew he had p chance of getting A.

Now the certainty of knowing which is which is higher w/ more info.

(30)

e) If we $\neq p$, how does this affect the decision rule?

It will tilt the decision to one of the coins (the one more likely to be picked) This will affect the observed distribution which Bob sees,

(31)

f) Find the value of p for which Bob will never guess he recieved the 1st coin, regardless of toss outcome.

If p is very small,

Like if p is 0, he will likly see 2 heads and declare his coin to be B, But there is a chance he will see only 1 head

$$\binom{3}{1}\left(\frac{2}{3}\right)^1\left(\frac{1}{3}\right)^2 \text{ if B} \quad + \binom{3}{0}\left(\frac{2}{3}\right)^0\left(\frac{1}{3}\right)^3$$

↑ prob $\frac{1}{3}$ that will be only 1 head w/ coin B

So this is not possible, because with only 3 tosses, still a fairly significant chance will declare coin B to be A

(32)

9) Now always guess 1st coin

I am guessing this is the same, but reverse

If $p = 1$ would always be A

But then chance to have 2, 3 heads

$$\binom{3}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^1 + \binom{3}{3}\left(\frac{1}{3}\right)^3 \qquad \text{(if is A)}$$

$$\frac{2}{9} + \frac{1}{9}$$

Same $\frac{1}{3}$ chance he will say its B and be wrong

So no good value for p

(33)

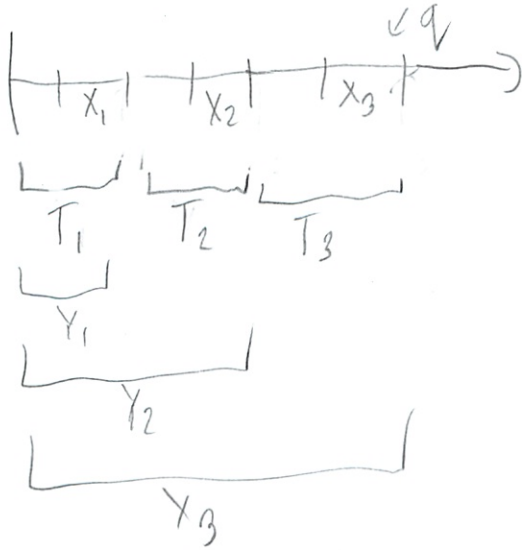
5. Consider Bernalli process $X_1, X_2, X_3 \ldots$
w/ unknown prob of sucess q
kth interarrival time T_k =

$$T_1 = Y_1 \quad T_k = Y_k - Y_{k-1} \quad k = 2, 3, \ldots$$

where Y_k is time of kth sucess
Explores estimation of q from observed inter-arrival time
Assume q is sampled from RV Q, which is
uniformly distributed $[0,1]$

Hint $\int_0^1 q^k (1-q)^m \, dq = \frac{k!\,m!}{(k+m+1)!}$

a) Compute PMF of T_1, $p_{T_1}(t_1)$



(-1)

(34)

b) Compute LSE of Q from 1st recording $T_1 = t_1$

So now have one reading

? Draw joint

?? how draw?

$$\hat{\theta}_{LMS} = E[\Theta \mid X = x]$$

$$= E[Q \mid T_1 = t_1]$$

$$= \int Q \underbrace{f_{Q|T}(q|x)}_{\text{? posterior dist}} dQ$$

?? so what is the posterior dist?

Want info on q

$$\int_0^1 q^k \, dq$$

(35)

So this probability of T_1 is time to first occival of Bernoulli (q)

But q is random $[0,1]$

- why type of problem is this?
- # trials to 1st sucess = geometric

$$(1-p)^{k-1} p \qquad k = 1$$

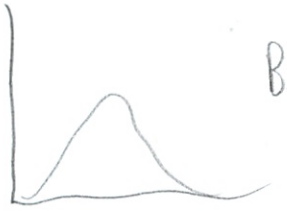
$$(1-p)^0 p$$

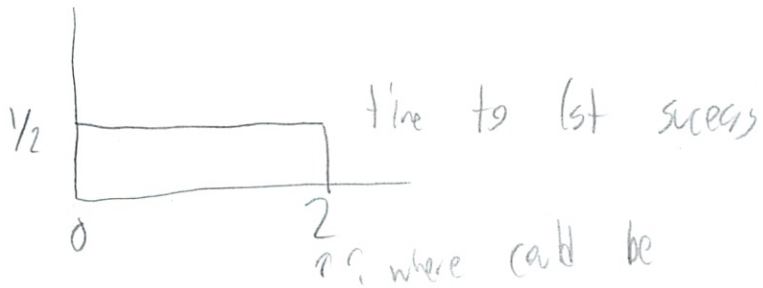
$$p$$

and $p = q$.

$$\frac{1}{b-a+1} = \frac{1}{1-0+1} = \frac{1}{2}$$

(-1)





?, where could be

(36)

c) Compute MAP estimate of Q given k recordings

$T_1 = t_1, \ldots, T_k = t_k$

Now have k estimates

Conduct potaior dist, find max value in it

$$f_{Q|T}(q \mid t = t_1, \ldots, t_k) = \frac{f_{T|Q}(t|q)\, f_Q(q)}{\int f_{T|Q}(t|q)\, f_Q(q)}$$

$f_{T|Q}(t|q)$

? just the geometric

$(1-p)^{k-1} p \quad k = 1$

p

$p = q$

So definitive value q

$f_Q(q) = \text{uniform} \quad \frac{1}{b-a+1} = \frac{1}{1-0+1} = \frac{1}{2}$

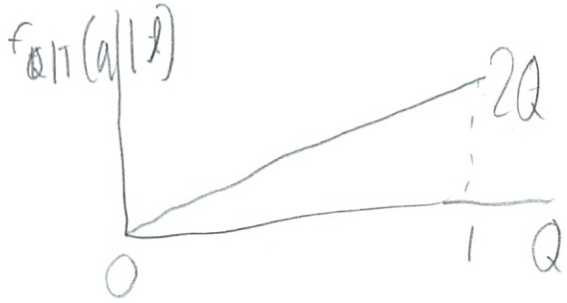
$\int_0^1 \frac{q}{2}\, dq = \frac{q^2}{4}\Big|_0^1 = \frac{1}{4}$

$= \frac{\frac{q}{2}}{\frac{1}{4}} = 2q$

?? this right

(37)

Need a single value - the max



? Max is $q = 1$ w/ prob 2

? can't be - prob > 1

(-1)

$f_{Q|T_1,\ldots T_n}(q|t_1,\ldots,t_n)$

$f_Q(q)\, f_{T_1,T_2\ldots,T_n|Q}(t_1,\ldots t_n|q)$

find q that maximizes this

* Given Q are ind. geometric

If not given Q, then not ind

Product in terms of $\underbrace{[1-q]^{-}}_{A(q)}\ \underbrace{(q)^{-}}_{B(q)}$

Sum in the exponent

take deriv of w/ respect to q

$$\frac{\partial A(q)}{\partial q} \quad \frac{\partial B(q)}{\partial q}$$

$$A(q)\frac{d}{dq}B(q) + B(q)\frac{d}{dq}A(q)$$

Set $=0$

- could be max, min

take 2nd deriv if min is $\ominus$

- well other way around

LSE/LMS

$$\hat{\theta}_{LMS}(x) = E[\theta | X = x]$$

↑measured X
got this

computing may not be easy
form of it is simple

When want to minimize the size of the error

Care about how big error is
- How far off are you

Not how often will you be wrong (MAP)

$\hat{\theta}_{LLMS} = aX + b$

best you can do is

$$= E[\theta] + \frac{cov(X,\theta)}{Var(X)}(X - E[X])$$ ← more intuitive

$$= E[\theta] + \rho \frac{\sigma_\theta}{\sigma_X}(X - E[X])$$ ← normalize

- amt above / below $E[\theta]$

(40)

For this part only assume q is sampled from RV Q which is uniform $[.5, 1]$

d) Find LLSE of 2nd interarrival time T_2 from Observed 1st arrival time $T_1 = t_1$

$$\hat{\Theta}_{LLMS} = E[\Theta] + \frac{Cov(X,\Theta)}{Var(X)}(X - E[X])$$

$$\hat{Q}_{LMS} = E[Q] + \frac{Cov(Q,T)}{Var(T)}(T - E[T]) \quad +0.5$$

$E[T]$ = geometric $= \frac{1}{p} = \frac{1}{q}$

$Var(T)$ = geometric $= \frac{1-p}{p^2} = \frac{1-q}{q^2}$

$E[Q]$ = Uniform $= \frac{a+b}{2} = \frac{3}{4}$ $\quad (-0.5)$

$Var(Q)$ = uniform $= \frac{(b-a)(b-a+2)}{12} = \frac{(1-.5)(1-.5+2)}{12} = \frac{.5 \cdot 2.5}{12} = \frac{5}{48}$

$Cov(Q,T) = E[QT] - E[Q] \cdot E[T]$ $\quad (-0.5)$

$E[QT] = E[E[QT \mid T=t]]$

$Q\,E[T \mid T=t]$

$Q \cdot \frac{1}{q}$

$Q \cdot \frac{1}{q} = 1$

(4)

$= E[1] = 1$

$\text{Cov}(Q, T) = 1 - \frac{3}{4} \cdot \frac{1}{q}$

? shall it have a variable

$\frac{1}{3/4}$

$= 0$ totally uncorrelated (-0.5)

$LLMS = \frac{3}{4} + \frac{0}{\text{rest does not matter}} \ldots$

$= \left(\frac{3}{4}\right)$

- Where did we base it on T_1 ?
- T_1
- This is for any T_{1a}

0.5/5

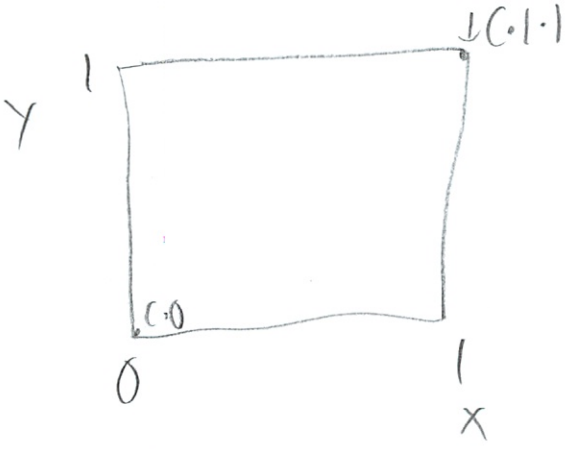
Have no clue if this is right

(42)

6. Joint PDF of X, Y defined

$$f_{X,Y}(x,y) = \begin{cases} cxy & \text{if } 0 < x \leq 1, 0 < y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

a) Find normalization constant c



$$\int_0^1 \int_0^1 cxy \, dy \, dx$$

$$\int_0^1 \frac{1}{2} cxy^2 dx \Big|_0^1$$

$$\int_0^1 \frac{cx}{2} dx$$

$$\frac{cx^2}{4} \Big|_0^1$$

$$\frac{c}{4} = \text{want} = 1$$

$$\frac{c}{4} = 1$$

$$c = 4$$

(43)

b) Compute the conditional expectation estimator of X based on observed value $Y = y$

Conditional expectation estimator = LMS estimator

$$\hat{\Theta}_{LMS} = E[\Theta \mid X = x]$$ general variables

$$= E[X \mid Y = y]$$ our variables

$$= \int x \underbrace{f_{X|Y}(x \mid y = y)}_{\text{Posterior def}} dx$$

$$f_{X|Y} = \frac{f_{X,Y}(x,y)}{f_Y(Y=y)} \quad \frac{4xy}{\text{& observed so w/ certainty} = 1}$$

$$= \int x \, 4xy \, dx$$

$$= \int_0^1 4x^2 y \, dx$$

$$\frac{4x^3 y}{3} \Big|_0^1$$

$\frac{4}{3}y$ ← So just plug in observed y

(44)

c) Is this estimate different what what you would have guessed before saw $Y = y$?

Yes, of carse because you new have additional information which you did not have before

Before $4xy$

After $\frac{4}{3}y$

(45)

d) Repeat for MAP

So just max of posterior definition $\frac{4_{x,y}}{1}$

What x will max $4_{x,y}$?

The largest possible $x \rightarrow x = (1)$

Yes this also provides slightly more information which we did not have before

Before $4_{x,y}$

After 4_y

It gives us an actual estimate, although it is a fairly inaccurate estimate. The new information is not very significant

Prob got all wrong

Problem Set 10 Solutions

1. **A financial parable.**

   (a) The bank becomes insolvent if the asset's gain $R \leq -5$ (i.e., it loses more than 5%). This probability is the CDF of R evaluated at -5. Since R is normally distributed, we can convert this CDF to be in terms of a standard normal random variable by subtracting away the mean and dividing by the standard deviation, and then look up the value in a standard normal CDF table.

   $$\begin{aligned} \mathbf{E}[R] &= 7, \\ \text{var}(R) &= 10^2 = 100, \\ \mathbf{P}(R \leq -5) &= \mathbf{P}\left(\frac{R-7}{10} \leq \frac{-5-7}{10}\right) = \Phi(-1.2) \approx 0.115. \end{aligned}$$

   Thus, by investing in just this one asset, the bank has a 11.5% chance of becoming insolvent.

   (b) If we model the R_i's as **independent** normal random variables, then their sum $R = (R_1 + \cdots + R_{20})/20$ is also a normal random variable (see Example 4.11 on page 214 of the text). Thus, we can calculate the mean and variance of this new R and proceed as in part (a). Note that since the random variables are assumed to be independent, the variance of their sum is just the sum of their individual variances.

   $$\begin{aligned} \mathbf{E}[R] &= (\mathbf{E}[R_1] + \cdots + \mathbf{E}[R_{20}])/20 = 7, \\ \text{var}(R) &= \frac{1}{20^2}(\text{var}(R_1) + \cdots + \text{var}(R_{20})) = \frac{20 \cdot 100}{400} = 5, \\ \mathbf{P}(R \leq -5) &= \mathbf{P}\left(\frac{R-7}{\sqrt{5}} \leq \frac{-5-7}{\sqrt{5}}\right) = \Phi(-5.367) \approx 0.0000000439 = 4.39 \cdot 10^{-8}. \end{aligned}$$

   Thus, by diversifying and assuming that the 20 assets have **independent** gains, the bank has seemingly decreased its probability of becoming insolvent to a palatable value.

   (c) Now, if the gains R_i are positively correlated, then we can no longer sum up the individual variances; we need to account for the covariance between pairs of random variables. The covariance is given by

   $$\text{cov}(R_i, R_j) = \rho(R_i, R_j)\sqrt{\text{var}(R_i)\text{var}(R_j)} = \frac{1}{2}\sqrt{10^2 \cdot 10^2} = 50.$$

   From page 220 in the text, we know that the variance in this case is

   $$\begin{aligned} \text{var}(R) &= \text{var}\left(\frac{1}{20}\sum_{i=1}^{20} R_i\right) = \frac{1}{400}\left(\sum_{i=1}^{20}\text{var}(R_i) + \sum_{\{(i,j)|i\neq j\}}\text{cov}(R_i, R_j)\right) \\ &= \frac{1}{400}(20 \cdot 100 + 380 \cdot 50) = 52.5. \end{aligned}$$

   Since we assume that $R = (R_1 + \cdots + R_{20})/20$ is still normal, we can again apply the same steps as in parts (a) and (b):

   $$\begin{aligned} \mathbf{E}[R] &= (\mathbf{E}[R_1] + \cdots + \mathbf{E}[R_{20}])/20 = 7, \\ \text{var}(R) &= 52.5, \\ \mathbf{P}(R \leq -5) &= \mathbf{P}\left(\frac{R-7}{\sqrt{52.5}} \leq \frac{-5-7}{\sqrt{52.5}}\right) = \Phi(-1.656) \approx 0.0488. \end{aligned}$$

Thus, by taking into account the positive correlation between the assets' gains, we are no longer as comfortable with the probability of insolvency as we thought we were in part (b).

2. Let M and N be the number of males and females, respectively, that cast a vote. We need to find $P(M > N)$, i.e., $P(M - N > 0)$. The central limit theorem does not apply directly to the random variable $M - N$. However, the central limit theorem implies that M and N are well approximated by normal random variables. So, $M - N$ is the difference of two independent approximately normal random variables. Since the difference of two normal random variables is itself normal, it follows that $M - N$ is approximately normal. The mean and variance of $M - N$ are found by

$$\mathbf{E}[M - N] = 300 \cdot 0.4 + 196 \cdot 0.5 = 120 - 98 = 22,$$
$$\text{var}(M - N) = \text{var}(M) + \text{var}(N) = 300 \cdot 0.4 \cdot 0.6 + 196 \cdot 0.5 \cdot 0.5 = 121.$$

Thus, the standard deviation of $M - N$ is 11. Let Z be a standard normal random variable. Using the central limit theorem approximation, we obtain

$$\begin{aligned} \mathbf{P}(M - N > 0) &= \mathbf{P}\Big(\frac{M - N - 22}{11} > -\frac{22}{11}\Big) \\ &\approx \mathbf{P}(Z \geq -2) \\ &= 0.9772. \end{aligned}$$

A slightly more refined estimate is obtained by expressing the event of interest as $\mathbf{P}(M - N \geq 1/2)$. We then have

$$\begin{aligned} \mathbf{P}(M - N > 1/2) &= \mathbf{P}\Big(\frac{M - N - 22}{11} \geq -\frac{21.5}{11}\Big) \\ &\approx \mathbf{P}(Z \geq -1.95) \\ &= 0.974. \end{aligned}$$

3. (a) Using the Central Limit Theorem, we obtain $\mathbf{P}(\frac{n}{2} - 10 \leq S_n \leq \frac{n}{2} + 10) \approx \Phi(\frac{20}{\sqrt{n}}) - \Phi(-\frac{20}{\sqrt{n}}) \to 0$ as $n \to \infty$.

   (b) The limit is 1, by the weak law of large numbers.

   (c) Using the Central Limit Theorem, we obtain $\mathbf{P}(\frac{n}{2} - \frac{\sqrt{n}}{2} \leq S_n \leq \frac{n}{2} + \frac{\sqrt{n}}{2}) \to \Phi(1) - \Phi(-1) = 0.6826$.

4. (a) Let C denote the coin that Bob received, so that $C = 1$ if Bob received the first coin, and $C = 2$ if Bob received the second coin. Then $\mathbf{P}(C = 1) = p$ and $\mathbf{P}(C = 2) = 1 - p$. Given C, the number of heads Y in 3 independent tosses is a binomial random variable.

   We can find the probability that Bob received the first coin given that he observed k heads using Bayes' rule.

$$\begin{aligned}
\mathbf{P}(C=1 \mid Y=k) &= \frac{\mathbf{P}(Y=k \mid C=1)\cdot \mathbf{P}(C=1)}{P(Y=k \mid C=1)\cdot \mathbf{P}(C=1)+\mathbf{P}(Y=k \mid C=2)\cdot \mathbf{P}(C=2)} \\
&= \frac{\binom{3}{k}\cdot (1/3)^k(2/3)^{3-k}p}{\binom{3}{k}\cdot (1/3)^k(2/3)^{3-k}\cdot p+\binom{3}{k}\cdot (2/3)^k(1/3)^{3-k}\cdot (1-p)} \\
&= \frac{2^{3-k}p}{2^{3-k}p+2^k(1-p)} = \frac{1}{1+\frac{1-p}{p}2^{2k-3}}
\end{aligned}$$

(b) We want to find k so that the following inequality holds.

$$\begin{aligned}
\mathbf{P}(C=1 \mid Y=k) &> p \\
\frac{2^{3-k}p}{2^{3-k}p+2^k(1-p)} &> p
\end{aligned}$$

Note that if $p=0$ or $p=1$, there is no value of k that satisfies the inequality. We now solve it for $0<p<1$:

$$\begin{aligned}
\frac{2^{3-k}}{2^{3-k}p+2^k(1-p)} &> 1 \\
2^{3-k} &> 2^{3-k}p+2^k(1-p) \\
2^{3-k}(1-p) &> 2^k(1-p) \\
2^{3-k} &> 2^k \\
2k &< 3 \\
k &< 3/2
\end{aligned}$$

For $0<p<1$, $k=0$ or $k=1$ the probability that Alice sent the first coin increases. The inequality does not depend on p, and so does not change when p increases. Intuitively, this makes sense: lower values of k increase Bob's belief he got the coin with lower probability of heads.

(c) Given that Bob observes k heads, Bob must decide on whether the first or second coin was used. To minimize the error, he should decide it is the first coin when $\mathbf{P}(C=1 \mid Y=k) \geq \mathbf{P}(C=2 \mid Y=k)$. Thus, we have the decision rule given by

$$\begin{aligned}
\mathbf{P}(C=1 \mid Y=k) &\geq \mathbf{P}(C=2 \mid Y=k) \\
\frac{2^{3-k}p}{2^{3-k}p+2^k(1-p)} &\geq \frac{2^k(1-p)}{2^{3-k}p+2^k(1-p)} \\
2^{3-k}p &\geq 2^k(1-p) \\
2^{2k-3} &\leq \frac{p}{1-p} \\
k &\leq \frac{3}{2}+\frac{1}{2}\log_2\frac{p}{1-p}
\end{aligned}$$

(d) i. If $p = 2/3$, the threshold in the rule above is equal to $\frac{3+\log_2 2}{2} = 2$. Therefore, Bob will decide that he received the first coin when he observes 0, 1 or 2 heads, and will decide that he received the second coin when he observes 3 heads.
We find the probability of a correct decision using the total probability law:

$$\begin{aligned}
\mathbf{P}(\text{Correct}) &= \mathbf{P}(\text{Correct} \mid C = 1) \cdot p + \mathbf{P}(\text{Correct} \mid C = 2) \cdot (1 - p) \\
&= \mathbf{P}(Y < 3 \mid C = 1) \cdot p + \mathbf{P}(Y = 3 \mid C = 2) \cdot (1 - p) \\
&= (1 - \mathbf{P}(Y = 3 \mid C = 1)) \cdot p + \mathbf{P}(Y = 3 \mid C = 2) \cdot (1 - p) \\
&= (1 - (1/3)^3)(2/3) + (2/3)^3(1/3) = 20/27 \approx .741
\end{aligned}$$

ii. In absence of any data, all Bob can do is decide he received the first coin with some probability q. Note that this rule includes the deterministic decisions that he received either the first coin ($q = 1$) or the second coin ($q = 0$).
In this case, the probability of correct decision is equal to

$$\begin{aligned}
\mathbf{P}(\text{Correct}) &= \mathbf{P}(\text{Correct} \mid C = 1) \cdot p + \mathbf{P}(\text{Correct} \mid C = 2) \cdot (1 - p) \\
&= qp + (1 - q)(1 - p) = 1 - p + q(2p - 1) = \frac{1 + q}{3}
\end{aligned}$$

Clearly, the probability of the correct decision is maximized (or the probability of error is minimized) when $q = 1$, i.e., when Bob deterministically decides he received the first coin. In this case, $\mathbf{P}(\text{Correct}) = 2/3 \approx .667$. Observing 3 coin tosses increases the probability of the correct decision by $2/27 \approx .074$.

(e) If p is increased, the threshold in the decision rule in part (c) goes up, i.e., the range of values of k for which Bob decides he received the first coin can only go up.

(f) Bob will never decide he received the first coin if the threshold in the rule above is below zero:

$$\begin{aligned}
\frac{3}{2} + \frac{1}{2} \log_2 \frac{p}{1-p} &< 0 \\
\log_2 \frac{p}{1-p} &< -3 \\
\frac{p}{1-p} &< \frac{1}{8} \\
p &< \frac{1}{9}
\end{aligned}$$

If $p < 1/9$, the prior probability of receiving the first coin is so low that no amount of evidence from 3 tosses of the coin will make Bob decide he received the first coin.

(g) Bob will always decide he received the first coin if the threshold in the rule above is equal to or above 3:

$$\begin{aligned}
\frac{3}{2} + \frac{1}{2} \log_2 \frac{p}{1-p} &\geq 3 \\
\log_2 \frac{p}{1-p} &\geq 3 \\
\frac{p}{1-p} &\geq 8 \\
p &\geq \frac{8}{9}
\end{aligned}$$

If $p \geq 8/9$, the prior probability of receiving the first coin is so high that no amount of evidence from 3 tosses of the coin will make Bob decide he received the second coin.

5. (a) Using the total probability theorem, we have

$$p_{T_1}(t) = \int_0^1 p_{T_1|Q}(t,q) f_Q(q) dq = \int_0^1 (1-q)^{t-1} q dq = \frac{1}{(t+1)t} \qquad \text{for } t = 1, 2, \ldots$$

(b) The least squares estimate coincides with the conditional expectation of Q given T_1, which is derived as

$$\begin{aligned}
\mathbf{E}[Q \mid T_1 = t] &= \int_0^1 p_{Q|T_1}(q \mid t) q dq \\
&= \int_0^1 \frac{p_{T_1|Q}(t \mid q) f_Q(q)}{p_{T_1}(t)} q dq \\
&= \int_0^1 t(t+1) q (1-q)^{t-1} q dq \\
&= \int_0^1 t(t+1) q^2 (1-q)^{t-1} dq \\
&= t(t+1) \frac{2(t-1)!}{(t+2)!} \\
&= \frac{2}{t+2}
\end{aligned}$$

(c) We write the posterior probability distribution of Q given $T_1 = t_1, \ldots, T_k = t_k$

$$\begin{aligned}
f_{Q|T_1,\ldots,T_k}(q \mid t_1, \ldots, t_k) &= \frac{f_Q(q) \prod_i^k P_{T_i}(T_i = t_i \mid Q = q)}{\int_0^1 f_Q(q) \prod_i^k P_{T_i}(T_i = t_i \mid Q = q) dq} \\
&= \frac{q^k (1-q)^{\sum_i^k t_i - k}}{c} \\
&= \frac{1}{c} q^k (1-q)^{\sum_i^k t_i - k},
\end{aligned}$$

where the denominator integrates out q so it could be viewed as a constant scalar c.
To maximize the above probability we set its derivative with respect to q to zero

$$k q^{k-1} (1-q)^{\sum_i^k t_i - k} - (\sum_i^k t_i - k) q^k (1-q)^{\sum_i^k t_i - k - 1} = 0,$$

or equivalently

$$k(1-q) - (\sum_i^k t_i - k) q = 0,$$

which yields the MAP estimate

$$\hat{q} = \frac{k}{\sum_{i=1}^k t_i}.$$

For this part only assume q is sampled from the random variable Q which is now uniformly distributed over $[0.5, 1]$

(d) The LLSE of T_1 given T_2 is

$$\hat{T}_2 = \mathbf{E}[T_2] + \frac{\text{cov}(T_1, T_2)}{\text{var}(T_1)}(T_1 - \mathbf{E}[T_1]),$$

where the coefficients are

$$\mathbf{E}[T_1] = \mathbf{E}[T_2] = \int_{0.5}^{1} f_Q(q)\mathbf{E}[T|Q=q]dq = \int_{0.5}^{1} 2 * 1/q dq = 2\ln 2,$$

and from the law of total variance

$$\begin{aligned}
\text{var}(T_1) = \text{var}(T_2) &= \mathbf{E}\left[\text{var}(T_1 \mid Q)\right] + \text{var}\left[\mathbf{E}(T_1 \mid Q)\right] \\
&= \mathbf{E}\left[\frac{1-Q}{Q^2}\right] + \text{var}\left[\frac{1}{Q}\right] \\
&= \mathbf{E}[1/Q^2] - \mathbf{E}[1/Q] + \mathbf{E}[1/Q^2] - \mathbf{E}[1/Q]^2 \\
&= \int_{0.5}^{2} f_Q(q)\frac{1}{q^2}dq - \int_{0.5}^{2} f_Q(q)\frac{1}{q}dq + \int_{0.5}^{2} f_Q(q)\frac{1}{q^2}dq - \left(\int_{0.5}^{2} f_Q(q)\frac{1}{q}dq\right)^2 \\
&= 2 - 2\ln 2 + 2 - (2\ln 2)^2 \\
&= 4 - 2\ln 2 - (2\ln 2)^2,
\end{aligned}$$

and their covariance

$$\begin{aligned}
\text{cov}(T_1, T_2) &= \mathbf{E}[T_1 T_2] - \mathbf{E}[T_1]\mathbf{E}[T_2] \\
&= \mathbf{E}\left[\mathbf{E}[T_1 T_2 \mid Q]\right] - \mathbf{E}[T_1]\mathbf{E}[T_2] \\
&= \mathbf{E}\left[\mathbf{E}[T_1 \mid Q]\mathbf{E}[T_2 \mid Q]\right] - \mathbf{E}[T_1]\mathbf{E}[T_2] \\
&= \mathbf{E}\left[1/Q^2\right] - \mathbf{E}[T_1]\mathbf{E}[T_2] \\
&= 2 - 4(\ln 2)^2
\end{aligned}$$

Therefore we have derived the linear least squares estimator

$$\hat{T}_2 = 2\ln 2 + \frac{2 - 4(\ln 2)^2}{4 - 2\ln 2 - (2\ln 2)^2}(T_1 - 2\ln 2) \approx 1.543 + 0.113T_1.$$

6. (a) To find the normalization constant c we integrate the joint PDF:

$$\int_0^1 \int_0^1 f_{X,Y}(x,y)\,dy\,dx = c\int_0^1 \int_0^1 xy\,dy\,dx = c\int_0^1 1/2x\,dx = c/4.$$

Therefore, $c = 4$.

(b) To construct the conditional expectation estimator, we need to find the conditional probability density.

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{4xy}{\int_0^1 4xy\,dx} = \frac{4xy}{2y} = 2x, \quad x \in (0,1]$$

Thus

$$\hat{x}_{\text{CE}}(y) = \mathbf{E}[X|Y=y] = \int_0^1 x \cdot 2x\,dx = 2/3.$$

(c) We first note that the conditional probability does not depend on y. Therefore, X and Y are independent, and whether or not we observe $Y = y$ does not affect the estimate in part (b). Another way to see this is to consider that if we do not observe y, we can compute the marginal $f_X(x) = \int_0^1 4xydy = 2x$ which is equal to the conditional density, and will therefore produce the same estimate.

(d) Since X and Y are independent, no estimator can make use of the observed value of Y to estimate X. The MAP estimator for X is equal to 1, regardless of what value y we observe, since the conditional (and the marginal) density is maximized at 1.

Recitation 23
December 2, 2010

1. **Example 9.1, page 463 in textbook**
   Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter θ is unknown. Assuming that Juliet was late by an amount x on their first date, find the ML estimate of θ based on the observation $X = x$.

2. **Example 9.4, page 464 in textbook**
   Estimate the mean μ and variance v of a normal distribution using n independent observations $X_1, \ldots, X_n$.

3. **Example 9.8, page 474 of textbook**
   We would like to estimate the fraction of voters supporting a particular candidate for office. We collect n independent sample voter responses $X_1, \ldots, X_n$, where X_i is viewed as a Bernoulli random variable, with $X_i = 1$ if the ith voter supports the candidate. We conducted a poll of 1200 people in North Carolina, and found that 684 were supporting the candidate. We would like to construct a 95% confidence interval for θ, the proportion of people who support the candidate. As we saw in lecture, using the central limit theorem, an (approximate) 95% confidence interval can be defined as

$$\hat{\Theta}^- = \hat{\Theta}_n - 1.96\sqrt{\frac{v}{n}}, \qquad \hat{\Theta}^+ = \hat{\Theta}_n + 1.96\sqrt{\frac{v}{n}}$$

   where $v = \mathrm{Var}(X_i)$, and $\hat{\Theta}_n = (X_1 + \ldots + X_n)/n$. Unfortunately, we don't know the value for v. Construct confidence intervals for θ using the following three ways of estimating or bounding the value for v (in each case simply assume that v is equal to the given estimate; note that this is a further approximation in cases (a) and (b)).

   (a)

$$\hat{S}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\Theta}_n)^2$$

   (b)

$$\hat{\Theta}_n(1 - \hat{\Theta}_n)$$

   (c) The most conservative upper bound for the variance.

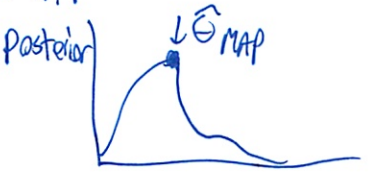
Recitation 23

12/2

- more stats
- highlights of stats
- infer obs based on unknows
- Bayesian – estimate Θ given X
  - Θ: unknown param; X: obs
  - ~~DoF~~ given $\underbrace{f_\Theta(\theta)}_{\text{prior}}$ $\underbrace{f_{X|\Theta}(x|\theta)}_{\text{conditional observation}}$
  - goal: get posterior w/ Bayes rule

$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)\, f_{X|\Theta}(x|\theta)}{f_X(x)}$$

  - lots of things to look at
    - $E[\Theta|X]$
    - LSE
    - MAP



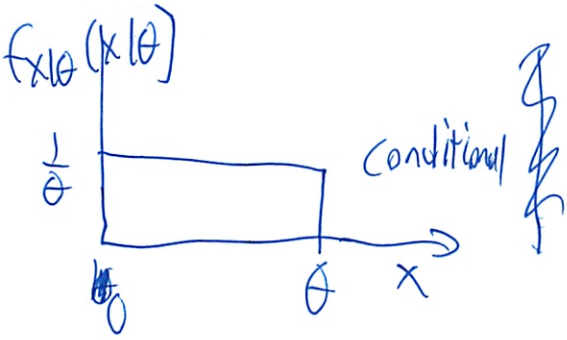
      - denom does not matter
      - $\arg\max_x f_\Theta(\theta) f_{X|\Theta}(x|\theta)$
      - can solve in closed form
      - w/o calculation

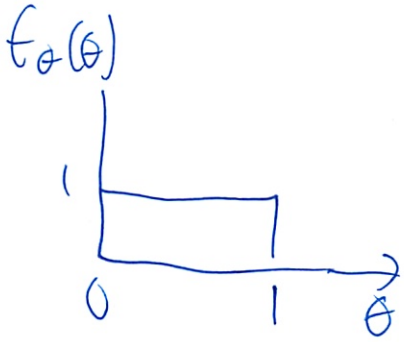
(2)

1. Romeo + Juliet are back again

Juliet late by amt X

$\sim$ uniform $[0, \theta)$





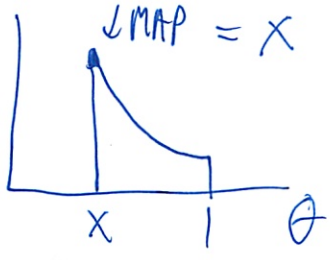
MAP rule

- maximize product of the two

$\max_\theta \frac{1}{\theta}$ for

- know $\theta > x$
  $\theta < 1$

So



- not so good
- check by calc var

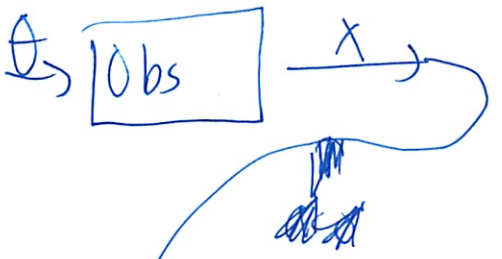
Classical Estimation

- given many prob models - one for each possible θ
- θ is unknown, not "random"

$f_x(x;\theta)$ ← one PDF for each value of θ

Observe x, try to find which θ it came from

One way → Max Likly-hood method (ML)



Consider all possible θ

for each θ, is some PDF for x

pick the one that is most likly

$f_x(x;\theta_1)$

$f_x(x;\theta_2)$

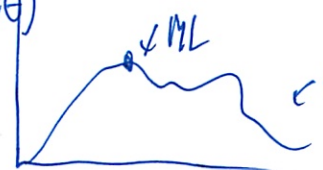
$\cancel{f_x(x;\theta_3)}$ ...

$f_x(x;\theta_m)$

max θ x in each θ → Max Likelyhood $\theta(x)$

Then for each $X = x$

$f_x(x;\theta)$



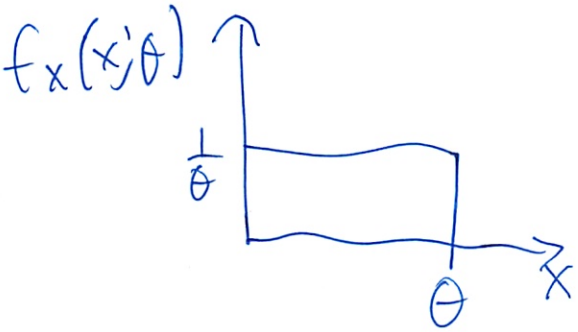
← find each value individually

(4)

R+J again

Give likelyhood for x for each value of θ

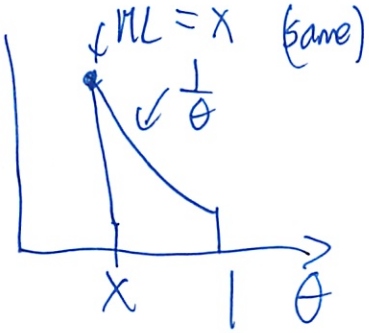
θ is ~~[illegible]~~ unknown except $0 \leq \theta \leq 1$



max over $x \leq \theta \leq 1$

↑ from problem definition

Plot likelyhood function for each θ



Is same

- since its a flat prior
- still 2 procedures are conceptually very different

(5)

2. multiple samples from normal – classical – ML

$X = X_1, \ldots X_n$ = iid Normal RV

$E[X_i] = \mu$

$Var(X_i) = V$

↑ we don't know var
so need to estimate it

$\theta = (\mu, V)$

↑ want, via likelyhood function

likelyhood fn $\rightarrow$ $f_n(X_1, \ldots, X_n; \mu, V) =$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi V}} \cdot e^{-(x_i-\mu)^2/2V}$$

$$= \frac{1}{(2\pi)^{n/2}} \cdot e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2V}$$

fancy algebra

$$= \frac{1}{(2\pi)^{n/2}} \cdot \frac{1}{V^{n/2}} \cdot e^{-n\left(s_n^2 + (m_n-\mu)^2\right)/2V}$$

where m_n = sample mean $= \sum_{i=1}^{n} X_i/n$

s_n^2 = avg of sum of deviations $= \frac{1}{n}\sum_{i=1}^{n}(X_i - m_n)^2$

~~exact~~ like sample var – but w/ sample mean

(6)

Can max log instead
- easier to max sums, instead of products

the "log likelihood"

$$\log f_X(x_1, \ldots, x_n \mid \mu, v) =$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(v) - \frac{n}{2v}s_n^2 - \frac{n(m_1-\mu)^2}{2v}$$

Tmax w/ respect to μ, v
can do deriv
but shortcut μ only in last term
and is positive

So only set last part = 0

$$\boxed{\hat{\mu} = m_1}$$

So max w/ respect to v of rest
- take deriv, = 0

$$\frac{\partial}{\partial v} = 0 \rightarrow -\frac{n}{2}\frac{1}{v} + \frac{n}{2}\frac{s_n^2}{v^2} = 0$$

$$\frac{s_n^2}{v} = 0$$

$$\boxed{\hat{v} = s_n^2}$$

(7) What we did

To estimate mean + var of normal, given n ind samples of it

We don't know if estimate is good

Check

$E[\hat{\mu}] = \mu$

↳ Same as true mean

called "unbiased"

$E[S_n^2] \neq V$

- but as $n \to \infty$

$E[S_n^2] \to V$

called "asymptotical unbiased-ness"

? ML has this in general

$\hat{\mu}_n \to \mu$ w/ prob 1

- special case WLLN

"Consitancy"

↳ becomes exact

$\hat{V}_n \to V$ close to WLLN

also "consitancy"

(8)

ML is one of the major methods for estimation

We had gotten estimates m_n, s_n^2 - but how confident are we?

- We introduce "confidence interval"
- again have multiple obs
  - but don't make point estimate $\hat{\theta}_n$
  - make lower + upper bounds of θ ("interval estimate")

$$[\theta_n^-, \theta_n^+]$$

- For every value of θ, we know θ is within the interval w/ high probability

$$\theta_n^- \leq \theta \leq \theta_n^+ \text{ w/ high prob}$$

↓ have as many intervals as have θ in classical

only 1 interval in Bayesian

θ_n^-, θ_n^+: RV - dist based on θ - estimator

$P(\theta_n^- \leq \theta \leq \theta_n^+)$ is a #

(9)

Say called $1-\alpha$ confidence interval

$$P(\theta_n^- \leq \theta \leq \theta_n^+) \geq 1-\alpha$$

~~Want estimate~~

Things to remember

1. θ_n^-, θ_n^+ are RVs (dist depends on θ)
2. Usually $\theta_n^- = \hat{\theta}_n - \beta$

   $\theta_n^+ = \hat{\theta}_n + \beta$ ← some reasonable estimator

Get end points w/ estimator
then go β left and right
adjust β till inequality holds

3. Often approx using the CLT

   θ_n^-, θ_n^+ calculated

8.2 again

$X = X_1, \ldots, X_n$

mean μ still unknown

var V now known

(10)

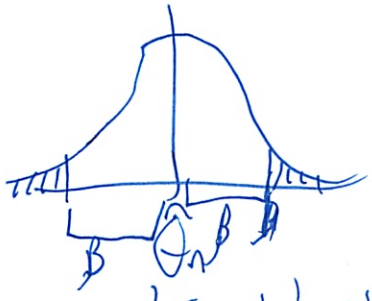
Adjust until threshold is met

- we know Var - so have idea of dist

$\hat{\theta}_n = \frac{X_1 + \ldots + X_n}{n}$ sample mean

Try CI $[\hat{\theta}_n - \beta, \hat{\theta}_n + \beta]$

Find β



↳ sample/estimated mean

$P(|\hat{\theta}_n - \theta| \leq \beta) \geq 1 - \alpha$

$P(|\theta_n - \theta| \leq \beta) \geq .95$

table

so $\beta \geq 1.96 \sqrt{\frac{V}{n}}$

↑ find smallest β that satisfies

$\left[\hat{\theta}_n - 1.96\sqrt{\frac{V}{n}}, \hat{\theta}_n + 1.96\sqrt{\frac{V}{n}}\right]$

(11)

But what if X_i not normal, but are iid

- Or if var is unknown

- So look at large sample → where sample mean ~ (almost) normal
- reasonably good approx
- but also need var - use nature of samples to get approx var

[1. Pretend sample mean is normal
2. Using approx var (estimated)

---

3. Pollster example

$X = X_1, \ldots, X_n$ Bernoulli

$E[X_i] = \theta$

Construct CI $[\hat{\theta}_n - B, \hat{\theta}_n + B]$

$\hat{\theta}_n$ = sample mean = fraction of 1s in sequence

Again (CLT) $B = 1.96 \sqrt{\frac{\hat{\sigma}_n^2}{n}}$ ← estimate of var (true var)

- of Bernoulli: $\theta(1-\theta)$

So $\hat{\theta}(1-\hat{\theta})$

or

2. Sample var
   - consistant estimator for normal or large # of samples
   - good estimator

$$\hat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_n)^2$$

w/ large # samples, not much different (>100 samples)

or

3. Use upper bound of var
   - for Bernalli $= \frac{1}{4} = \hat{V}_n$
   - But will over estimate
   - conservative CI

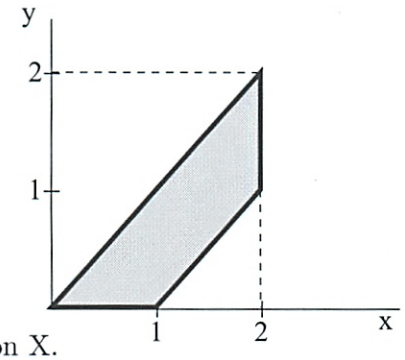
- Same type of calc in lecture 19 except here estimate var

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Electrical Engineering & Computer Science
6.041/6.431: Probabilistic Systems Analysis
(Fall 2010)

Tutorial 11

1. Continuous random variables X and Y have a joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 2/3 & \text{if } (x,y) \text{ belongs to the closed shaded region} \\ 0 & \text{otherwise} \end{cases}$$



We want to estimate Y based on X.

   (a) Find the LMS estimator $g(X)$ of Y.

   (b) Calculate the conditional mean squared error $\mathbf{E}\left[(Y - g(X))^2 \mid X = x\right]$.

   (c) Calculate the mean squared error $\mathbf{E}\left[(Y - g(X))^2\right]$. Is it the same as $\mathbf{E}\left[\text{var}(Y|X)\right]$?

   (d) Derive $L(X)$, the linear LMS estimator of Y based on X.

   (e) How do you expect the mean squared error of $L(X)$ to compare to that of $g(X)$?

   (f) What problem do you expect to encounter, if any, if you try to find the MAP estimator for Y based on observations of X.

2. Consider a noisy channel over which you send messages consisting of 0s and 1s to your friend. It is known that the channel independently flips each bit sent with some fixed probability p; however the value of p is unknown. You decide to conduct some experiments to estimate p and seek your friend's help. Your friend, cheeky as she is, insists that you send her messages consisting of three bits each (which you will both agree upon in advance); for each message, she will only tell you the total number of bits in that message that were flipped. Let X denote the number of bits flipped in a particular three-bit message.

   (a) Find the PMF of X.

   (b) Derive the ML estimator for p based on $X_1, \ldots, X_n$, the numbers of bits flipped in the first n three-bit messages.

   (c) Is the ML estimator unbiased?

   (d) Is the ML estimator consistent?

   (e) You send $n = 100$ three-bit messages and find that the total number of bits flipped is 20. Construct a 95% confidence interval for p. If necessary, you may use a conservative bound on the variance of the number of bits flipped.

   (f) What are some other ways to estimate the variance. How do you expect your confidence interval to change with different estimates of the variance.

Tutorial 11

12/3

Bayesian Inference

have θ → unknown, can't observe directly

X = observations

Want to make inference about θ

θ → [measurement] → X → [inference] → $\hat{\theta} = g(x)$

θ: ↑RV $P_\theta(\theta)$

measurement: ↑$P_{X|\theta}(x=x|\theta)$

assignment in P-set

first step: get posterior prob

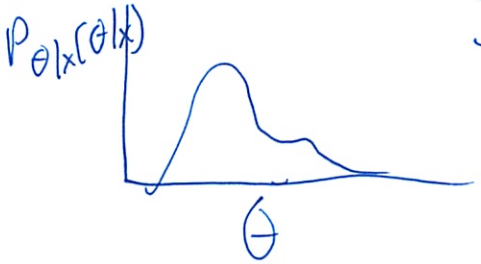
$$P_{\theta|X}(\theta|x) = \frac{P_{X|\theta}(x|\theta)\, P_\theta(\theta)}{P_X(x)}$$

Bayes Rule

? pr $\sum_\theta P_{X|\theta}\, P_\theta(\theta)$

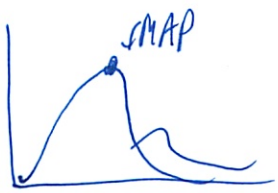
Now have

- a distribution
- want a single # estimate



error = result − actual

$\hat{\theta}$ MAP $= \arg\max_\theta P_{\theta|X}(\theta|x)$



minimum probability of error

(2)

Find by taking deriv, set = to 0, solve for θ

LMS/CEE

- minimizes mean square error

$$E[(\theta - \hat{\theta})^2]$$

$$\hat{\theta}_{LMS} = E[\theta | X]$$ ← often hard to find, complicated

LLMS

restrict estimator to linear to simplify

$$\hat{\theta} = E[\theta] + \frac{Cov(X,\theta)}{Var(\theta)}(X - E[X])$$

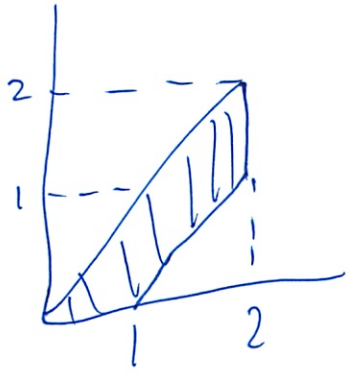
will have higher mean square error

or =

MMSE when posterior PDF is linear

#1 Given this joint PDF

Uniform in shaded region



(3)

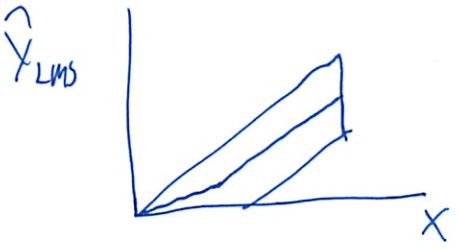
a) $\hat{Y}_{LMS} = E[Y|X]$

$Y|X$ is 

if uniform, $E[\,]$ is midpoint of line

$$= \begin{cases} \frac{X}{2} & \text{if } X \in [0,1) \\ \frac{X+X-1}{2} = X - \frac{1}{2} & \text{if } X \in [1,2] \end{cases}$$

Can plot estimator for each value of X



b) Now the error that results from using this estimator

$$E[(Y-g(X))^2 | X=x] = E[(Y-E[Y|X])^2 | X=x] =$$

- is the conditional var in general

$$= Var(Y|X=x)$$

$$= \begin{cases} \frac{x^2}{12} & \text{if } x \in [0,1] \\ \frac{1}{12} & \text{if } x \in [1,2] \end{cases}$$

(4)

c) Now just mean squared error, no conditioning

$$E[(Y-g(X))^2] \stackrel{?}{=} E[\text{var}(Y|X)]$$

? have

use iterate expectation

statement is true

$$= E[E[(Y-g(X))^2 | X]$$

substitute from B

$$= E[\text{var}(Y|X)]$$

? in general this one above was if have a specific $X=x$

$$= \int_0^2 \text{var}(Y|X=x) \underbrace{f_X(x)}\, dy$$

need to find by integrating over the ys

bounds are tricky

$$f_X(x) = \int_0^2 f_{X,Y}(x,y)\, dy$$

must break bounds down

$$= \begin{cases} \int_0^x \frac{2}{3}\, dy & x \in [0,1) \\ \int_{x-1}^x \frac{2}{3}\, dy & x \in [1,2] \end{cases}$$

(5)

$$= \begin{cases} \frac{2}{3}x & x \in [0,1] \\ \frac{2}{3} & x \in [2,3] \end{cases}$$

$$= \int_0^1 \frac{x^2}{12} \cdot \frac{2x}{3}\,dx + \int_1^2 \frac{1}{12} \cdot \frac{2}{3}\,dx$$

$$= \frac{5}{72}$$

d) LLMS

- just plug in formula

$$\hat{Y} = L(X) = E[Y] + \frac{Cov(X,Y)}{Var(Y)}(X - E[X])$$

- p-set 4,5 style integrations and such
- skip

e) Compare LLMS to LMS

- less accurate
- LLMS error $\leq$ LMS error
  - LLMS error: can calculate w/ same process as c
  - LMS error: part c
- also in book

$$\sigma_Y^2(1-\rho^2)$$

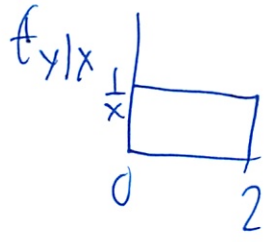
(6)

f) MAP for y based on X

$$\hat{y}_{MAP} = \arg\max_{y} f_{Y|X}(y|x)$$

↑ will be uniform for any particular x

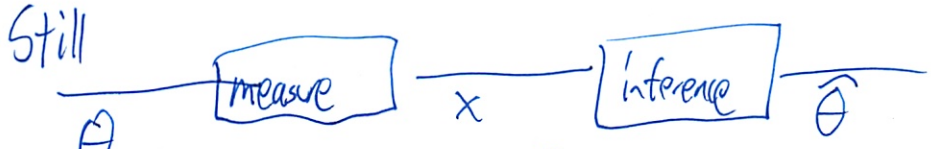
But how find the max of something uniform?



Pick arbitrarily

No help

Classical Estimation



↑ just an RV, no prior

$P(x;\theta)$ ↑ $P_X(x)$ for each θ

Can't calc posterior

Can't do Bayes

So Max Likly hood → $\hat{\theta}_n = \arg\max P(x_1, x_2, \ldots, x_n; \theta)$

(7)

Same ans as MAP in some cases

MAP $\underset{\theta}{\operatorname{argmax}}\ P(\theta \mid X_1, \ldots, X_n)$

$= \underset{\theta}{\operatorname{argmax}}\ P(X_1, X_2, \ldots X_n \mid \theta)\, P_\theta(\theta)$

($P(X_1, X_2, \ldots X_n \mid \theta)$: does not matter)

→ when prior is uniform

$\hat{\theta}_{ML} = \hat{\theta}_{MAP}$

---

2. PS 10 was hyp testing; Bayesian Inf

This is classical; and find p — not just test hyp

a) PMF of X

~ binomial (p)

$$P_X(k; p) = \begin{cases} \binom{3}{k} p^k (1-p)^{3-k} & k = 0, 1, 2, 3 \\ 0 & \text{else} \end{cases}$$

b) $X_1, \ldots, X_n$ = # of bits flipped in 1st n messages

$$\hat{p}_{ML} = \underset{p}{\operatorname{argmax}}\ P(X_1, X_2, \ldots, X_n; p)$$

- messages are ind

(8)

So can replace joint with a product when i.i.d

$$= \arg\max_p \prod_{i=1}^{n} P(X_i; p)$$

$X_i = k_i$

$$= \arg\max_p \prod_{i=1}^{n} \binom{3}{k_i} p^{k_i} (1-p)^{3-k_i}$$

? same as part a, ✱

take deriv, set $= 0$

- really hard to do w/ product
  - ↳ chain rule
- take log to get rid of product
  - how is this ok?
    - if strictly monotonic increasing → will be increasing ↳ 1 to 1 map

$$= \arg\max_p \log\left(\prod_{i=1}^{n} P(X_i; p)\right)$$

log of product = sum of products

$$= \arg\max_p \sum_{i=1}^{n} \log \binom{3}{k_i} p^{k_i} (1-p)^{3-k_i}$$

turns into a sum → again log of product = sum of products

$$= \arg\max_p \sum_{i=1}^{n} \log \binom{3}{k_i} + \sum_{i=1}^{n} k_i \log p + \sum_{i=1}^{n} (3-k_i) \log(1-p)$$

($\log p^{k_i}$; $\log(1-p)^{(3-k_i)}$)

derivative

simplified

$\log a^x = x \log a$

$$= 0 + \sum_{i=1}^{n} \frac{k_i}{p} - \sum_{i=1}^{n} \frac{(3-k_i)}{1-p}$$

set $= 0$

(a)

Simplify

$$\hat{p}_{n_{ML}} = \frac{1}{3n} \sum_{i=1}^{n} k_i$$

↑ "makes sense"
if 20 out of 300 flipped
$p \approx \frac{20}{300}$

$$= \frac{1}{3n} \sum_{i=1}^{n} X_i$$

↑ since valid for all ks (little x)

c) Is estimator unbiased

- Basian → calc MBE error, p error (to check if good)
- Classical → $E[\hat{p}_n] = p$ ← our guess for

$$= E\left[\frac{1}{3n} \sum_{i=1}^{n} X_i\right]$$

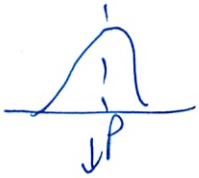
$$= \frac{\sum_{i=1}^{n} E[X_i]}{3n}$$

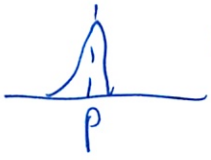
$$= \frac{3pn}{3n} = p$$ ✓ yes

(10)

d) Consistent: $\hat{p}_n \xrightarrow{\text{in prob}} p$

Converges in prob $\lim_{n\to\infty} P(|\hat{p}_n - p| \geq 0) = 0 \quad \forall \epsilon > 0$





ϵ gets more + more centered around p

$\hat{p}_n = \underbrace{\sum_{i=1}^{n} X_i}_{3_n} \to p$

Sample mean $\frac{\sum X_i}{n} \to 3p$

Due to WLLN

e) Confidence Interval

- report interval, not just estimate

$E[\hat{p}_n^-, \hat{p}_n^+]$

$P(\hat{p}_n^- \leq p \leq \hat{p}_n^+) \geq \underbrace{1-\chi}_{\text{Confidence interval} = 95\%}$ given

(11)

Start w/ finding $\hat{p}_n^-$, $\hat{p}_n^+$

So start w/ $\hat{p}_n$

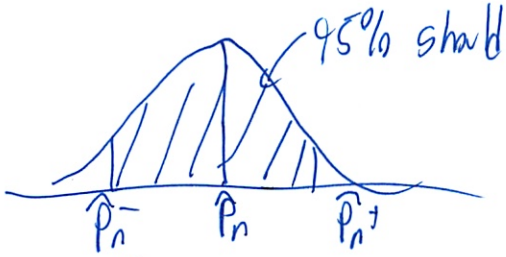
$$\hat{p}_n = \frac{\sum_{n=1}^{3n} Y_i}{m}$$

Y_i ← sum of 1 bits

m ← total # bits

← sum up all bits, not just all messages

$\hat{p}$ ≈ normal for large n



← need to find actual bounds

$$P\left(-z \le \frac{\hat{p}_n - p}{\sigma_{\hat{p}_n}} \le z\right) = .95$$

$z = ?$

Look at st normal so CDF is .025

$z = 1.96$ ← going backwards

~~$\hat{p}_n^+ = \hat{p}_n + 1.96$~~

~~$\hat{p}_n^- = \hat{p}_n - 1.96$~~

(12) but need to get it out of "normal form" - denormalize

$$P\left(-1.96 \le \frac{\bar{P}_n - p}{\sigma} \le 1.96\right) \ge .095$$

$$P\left(\hat{P}_n - 1.96\sqrt{\frac{V}{n}} \le p \le \hat{P}_n + 1.96\sqrt{\frac{V}{n}} \leftarrow \text{var}\right)$$

↑ total #

If don't know var, guess!

Conservative $\frac{1}{4}$ for bernoulli

Or sample var $\hat{P}_n(1-\hat{P}_n)$

$$\hat{P}_n^{\text{bd}} = \hat{P}_n + 1.96\sqrt{\frac{1/4}{300}}$$

$= a \#$