12/6

LECTURE 24

- **Reference:** Section 9.3
- Course VI Underground Guide Evaluations
  https://sixweb.mit.edu/student/evaluate/6.041-f2010
  https://sixweb.mit.edu/student/evaluate/6.431-f2010

Will still be on final

Note statement conflicted in review section

Outline

More classical stats

- Review
  - Maximum likelihood estimation
  - Confidence intervals
- Linear regression ← Some ways to mis use
- Binary hypothesis testing
  - Types of error
  - Likelihood ratio test (LRT)

Review

- **Maximum likelihood estimation**
  - Have model with unknown parameters: $X \sim p_X(x;\theta)$
  - Pick θ that "makes data most likely"

$$\max_\theta p_X(x;\theta)$$

θ is constant we do not know

Got x

Look at max liklyhood

Want value of θ where x is most liky to happen

  - Compare to Bayesian MAP estimation:

$$\max_\theta p_{\Theta|X}(\theta \mid x) \quad \text{or} \quad \max_\theta \frac{p_{X|\Theta}(x|\theta)p_\Theta(\theta)}{p_Y(y)}$$

θ is RV

Find value of θ, which this is largest

So most liky θ

Same if prior is constant

Denominator does not matter

- **Sample mean estimate of $\theta = \mathrm{E}[X]$**

$$\hat{\Theta}_n = (X_1 + \cdots + X_n)/n$$

Collect samples - find avg

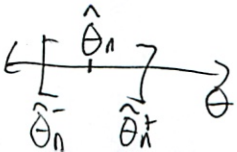
- $1-\alpha$ **confidence interval**

How far away do you think it is?

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

- confidence interval for sample mean
  - let z be s.t. $\Phi(z) = 1 - \alpha/2$

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



← So θ inside interval w/ prob $1-\alpha$
Or $1-\alpha$ points will fall within

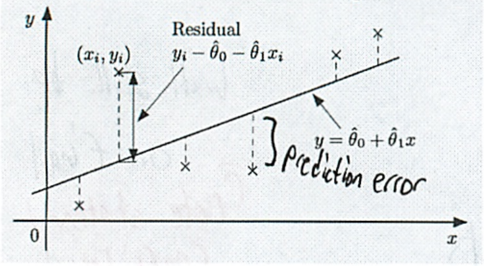
$P(2.1 \le \theta \le 3.9) \ge .95$ ← False w/ little θ - it's a #
So leave it as RVs - the interval - just that interval falls around θ

← ie 95% of experiments will fall inside interval. Can use CLT to approx ⊕

$P\left(\frac{\sqrt{n}}{\sigma}(\Theta_n - \underbrace{E[X]}_{\boxed{\theta}}) \geq z\right) \approx 2(1-\Phi(z))$

Regression

(Remember from high school)



half of what staticians do - most common model

Do my Xs provide useful info of predicting Ys

X = HS GPA

Y = MIT GPA

Find line that best fits data

- Data: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Model: $y \approx \theta_0 + \theta_1 x$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2 \qquad (*)$$

← Sum of square of error as small as possible

- One interpretation: Relation to Prob
  $Y_i = \theta_0 + \theta_1 x_i + W_i, \quad W_i \sim N(0, \sigma^2)$, i.i.d.
  - Likelihood function $f_{X,Y|\theta}(x, y; \theta)$ is:

$$c \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2\right\}$$

  - Take logs, same as (*)
  - Least sq. ↔ pretend W_i i.i.d. normal

- same as ML w/ some assumptions

Random element ~normal

0 mean

Same var σ^2

try to make as large as possible do exponential then as small as possible

density function of Y vector

$$f_{Y_1, Y_2, \ldots, Y_n}(y_1, y_2, \ldots, y_n; \theta_0, \theta_1) = \prod_{i=1}^{n} f_{Y_i}(y_i; \theta)$$

Linear regression

- **Model** $y \approx \theta_0 + \theta_1 x$

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

derivitive will be linear

2 Unknowns θ_0, θ_1

- **Solution** (set derivatives to zero):

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}, \qquad \overline{y} = \frac{y_1 + \cdots + y_n}{n}$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) / n}{\sum_{i=1}^{n} (x_i - \overline{x})^2 / n}$$

natural estimate of var

$$\hat{\theta}_0 = \overline{y} - \hat{\theta}_1 \overline{x}$$

Solution intuitive

$Y = \theta_0 + \theta_1 X + W$

$E[Y] = \theta_0 + \theta_1 E[X]$

$\overline{Y} = \theta_0 + \theta_1 \overline{x} \rightarrow \hat{\theta}_0 = \overline{Y} - \theta_1 \overline{x}$

suppose $E[X] = E[Y] = 0$

$E[W] = 0$

- **Interpretation** of the form of the solution
  - Assume a model $Y = \theta_0 + \theta_1 X + W$
    W independent of X and Y, with zero mean
  - Check that

$$\theta_1 = \underbrace{\text{cov}(X, Y)}_{Var(X)} = \frac{\mathbf{E}\big[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])\big]}{\mathbf{E}\big[(X - \mathbf{E}[X])^2\big]}$$

corrections

  - Solution formula for $\hat{\theta}_1$ is a natural estimate of the covariance

~~$E[Y^2] = \theta_0 + \theta_1 E[XY]$~~

$\hat{\theta}_1 = \frac{E[XY]}{Var(Y)} = \frac{Cov(X,Y)}{Var(Y)}$

$E[XY] = \cancel{\theta_0} + \theta_1 E[X^2] + \cancel{E[XW]}$

The world of linear regression

- **Multiple linear regression:**
  - **data:** (x_i, x'_i, x''_i, y_i), $i = 1, \ldots, n$ if more than 1 variable
  - **model:** $y \approx \theta_0 + \theta x + \theta' x' + \theta'' x''$
  - **formulation:**

$$\min_{\theta,\theta',\theta''} \sum_{i=1}^{n} (y_i - \theta_0 - \theta x_i - \theta' x'_i - \theta'' x''_i)^2$$

X_1= HS GPA

X_2= Family Income

X_3= SAT score

explanitory variables

Y = RIT GPA

- **Choosing the right variables**
  - model $y \approx \theta_0 + \theta_1 h(x)$ linear
    e.g., $y \approx \theta_0 + \theta_1 x^2$ quadratic
  - work with data points $(y_i, h(x))$
  - formulation:

again set derivs =0 and solve

must make a choice

$$\min_{\theta} \sum_{i=1}^{n} (\underbrace{y_i}_{\text{outcome}} - \underbrace{\theta_0 - \theta_1 h_1(x_i)}_{\text{prediction}})^2$$

- just diff set of explanitory variables

The world of regression (ctd.)

- **In practice,** one also reports
  - Confidence intervals for the θ_i - not explaining, but established methodology
  - "Standard error" (estimate of σ) - Some noise will always be there
  - R^2, a measure of "explanatory power"
    "factor explains 20% of variation"

$$1 - \frac{Var(Y|X)}{Var(Y)}$$

← var in Ys if know X, residual - how much left

← var in Ys

- **Some common concerns**
  - Heteroskedasticity
  - Multicollinearity
  - Sometimes misused to conclude causal relations - biggest mistake
  - etc.

lots of issues to watch out for

Heteroskedasticity

var gets bigger as x gets larger

- strongly affected by later data, less affected by low x data

Multicollinearity

$Y = \theta + \theta_1(HSGPA) + \theta_2^2(HSGPA)$

its the same thing!
or its closly related
both can explain it

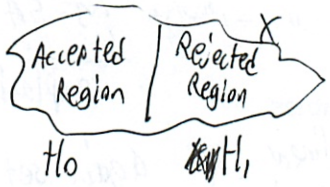
Fitting Line just tells us closely related
- causality can go either way or none of the above - only association

Binary hypothesis testing as well as next time

- Binary θ; new terminology:
  - **null hypothesis** H_0: default
    $X \sim p_X(x; H_0)$ [or $f_X(x; H_0)$]
  - **alternative hypothesis** H_1: new
    $X \sim p_X(x; H_1)$ [or $f_X(x; H_1)$]

two diff dist of X
-did x come from 1 dist or the other

- Partition the space of possible data vectors
  Rejection region R:
  reject H_0 iff data $\in R$



depends where Xs fall

- Types of errors:
  - **Type I (false rejection**, false alarm):
    H_0 true, but rejected
    $$\alpha(R) = \mathrm{P}(X \in R; H_0)$$
  - **Type II (false acceptance,** missed detection):
    H_0 false, but accepted
    $$\beta(R) = \mathrm{P}(X \notin R; H_1)$$

(airline industry)

if shift size of region, ~~only make~~
tradeoff of errors

Likelihood ratio test (LRT)

- Bayesian case (MAP rule): choose H_1 if:
  $\mathrm{P}(H_1 \mid X = x) > \mathrm{P}(H_0 \mid X = x)$
  or
  $$\frac{\mathrm{P}(X = x \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(X = x)} > \frac{\mathrm{P}(X = x \mid H_0)\mathrm{P}(H_0)}{\mathrm{P}(X = x)}$$
  or
  $$\frac{\mathrm{P}(X = x \mid H_1)}{\mathrm{P}(X = x \mid H_0)} > \frac{\mathrm{P}(H_1)}{\mathrm{P}(H_0)}$$
  (likelihood ratio test)

Bayes rule
easy to see which error is bigger

decision procedure
if odds in favor, choose H_1

- Nonbayesian version: choose H_1 if
  $$\frac{\mathrm{P}(X = x; H_1)}{\mathrm{P}(X = x; H_0)} > \xi \quad \text{(discrete case)}$$
  $$\frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi \qquad \text{(continuous case)}$$

Under H_1, how likely to observe this x

~~a~~ ~~that~~ with X observed - is it more likely to be produced by H_0 or H_1

- threshold ξ trades off the two types of error
  - choose ξ so that P(reject H_0; H_0) $= \alpha$
    (e.g., $\alpha = 0.05$)

(a) Find the ML estimates of the linear model parameters.

(b) Find the ML estimates of the quadratic model parameters.

Note: You may use the regression formulas and the connection with ML described in pages 478-479 of the text. However, the regression material is outside the scope of the final.

The figure below shows the data points (x_i, y_i), $i = 1, \ldots, 5$, the estimated linear model

$$y = 40.53x - 65.86,$$

and the estimated quadratic model

$$y = 4.09x^2 - 3.07.$$

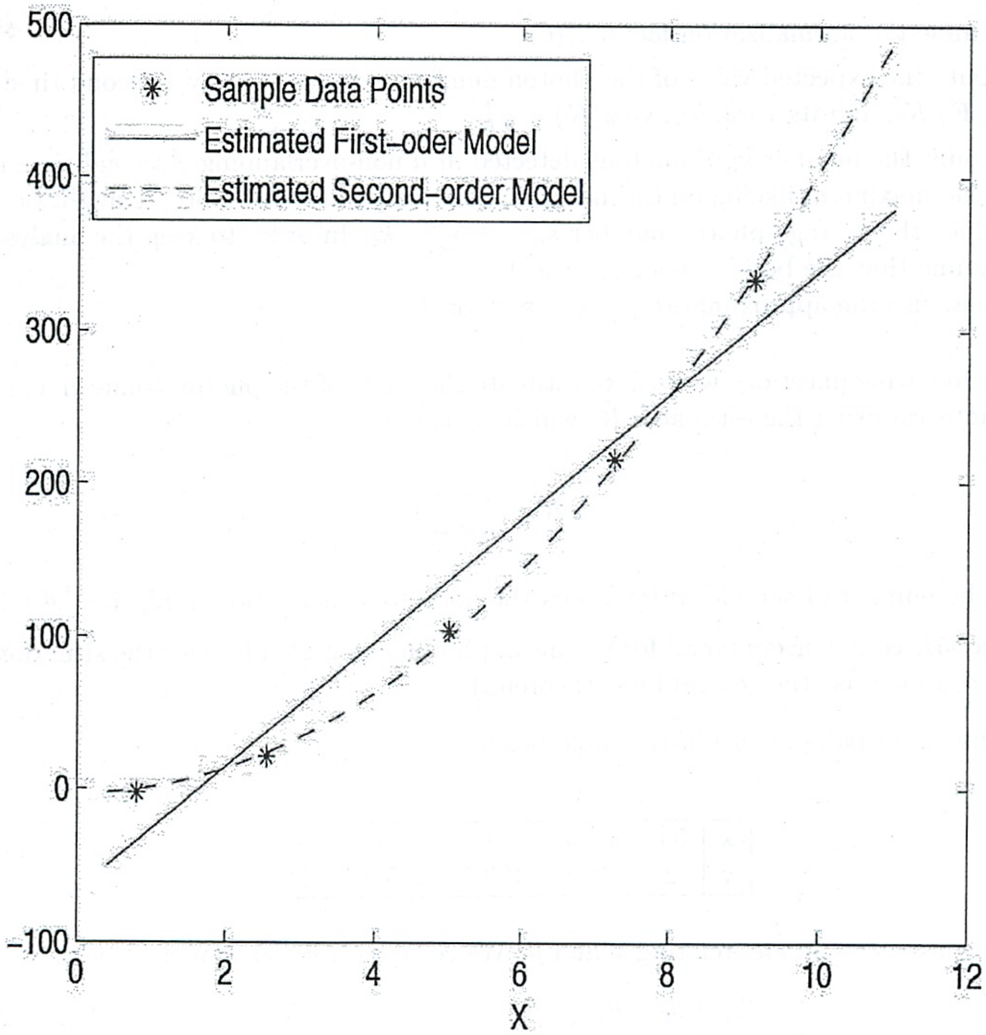


Figure 1: Regression Plot

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

Recitation 24
December 7, 2010

1. A blackbody at temperature θ radiates photons of all wavelengths, described by its characteristic spectrum. This problem will have you estimate θ, which is fixed but unknown. The PMF for the number of photons K in a given wavelength range and a fixed very short time interval is given by,

$$p_K(k;\theta) = \frac{1}{Z(\theta)} e^{-k/\theta}, k = 0, 1, 2, ...$$

$Z(\theta)$ is a normalization factor for the probability distribution (the physicists call it the partition function). You are given the task of determining the temperature of the body to two significant digits by photon counting in non-overlapping time intervals of duration one second. The photon emissions in non-overlapping time intervals are statistically independent from each other.

   (a) Determine the normalization factor $Z(\theta)$.

   (b) Compute the expected value of the photon number measured in any 1 second time interval, $\mu_K = \mathrm{E}_\theta[K]$, and its variance, $\mathrm{var}_\theta(K) = \sigma_K^2$.

   (c) You count the number k_i of photons detected in n non-overlapping 1 second time intervals. Find the maximum likelihood estimator, $\hat{\theta}_n$, for temperature θ. Note, it might be useful to introduce the average photon number $s_n = \frac{1}{n}\sum_{i=1}^{n} k_i$. In order to keep the analysis simple we assume that the body is hot, i.e. $\theta \gg 1$.
   You may use the approximation: $\frac{1}{e^{1/\theta}-1} \approx \theta$ for $\theta \gg 1$.

   In the following questions we wish to estimate the mean of the photon count in a one second time interval using the estimator $\hat{K}$, which is given by,

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{n} K_i.$$

   (d) Find the number of samples n for which the noise to signal ratio for $\hat{K}$, (i.e.,$\frac{\sigma_{\hat{K}}}{\mu_{\hat{K}}}$), is 0.01.

   (e) Find a 95% confidence interval for the mean photon count estimate for the situation in part (d). (You may use the central limit theorem.)

2. Given the five data pairs (x_i,y_i) in the table below,

x	0.8	2.5	5	7.3	9.1
y	-2.3	20.9	103.5	215.8	334

we want to construct a model relating x and y. We consider a linear model

$$Y_i = \theta_0 + \theta_1 x_i + W_i, \qquad i = 1, \ldots, 5,$$

and a quadratic model

$$Y_i = \beta_0 + \beta_1 x_i^2 + V_i, \qquad i = 1, \ldots, 5.$$

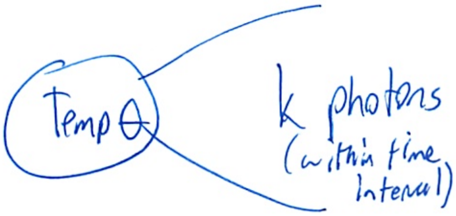
where W_i and V_i represent additive noise terms, modeled by independent normal random variables with mean zero and variance σ_1^2 and σ_2^2, respectively.

Recitation 24

12/7

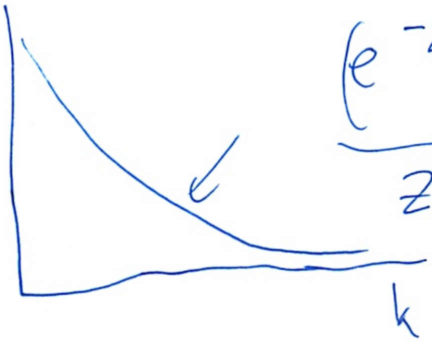
- Inference/chap 9 review

1. Material



Measure photons to estimate temp

$p_K(k;\theta)$



$$\frac{\left(e^{-1/\theta}\right)^k}{Z(\theta)}$$

for given value of θ

↳ changes param of geometric ↗ but diff pick for each θ

a) What is $Z(\theta)$?

$$1 = \frac{1}{Z(\theta)} \sum_{k=0}^{\infty} \left(e^{-1/\theta}\right)^k$$

$$= \frac{1}{Z(\theta)} \frac{1}{1-e^{-1/\theta}}$$

$$Z(\theta) = \frac{1}{1-e^{-1/\theta}}$$

②

b) Find mean + var $\mu_k = E_\theta[k]$ $\sigma_k^2 = var_\theta(k)$

- is a geometric
- but is shifted to right by 1

↳ param $p = (1-e^{-1/\theta})$

$$\mu_k = \frac{1}{p} - 1$$

$$= \frac{1}{1-{}^{-1/\theta}} - 1$$

$$= \frac{1}{e^{1/\theta}-1}$$

$$Var_\theta(k) = \sigma_k^2 = \underbrace{\frac{1-p}{p^2}}_{\text{Var of Geometric}}$$

shift does not matter

$$= \frac{e^{-1/\theta}}{(1-e^{-1/\theta})^2}$$

.... (algebra)

$$= \mu_k^2 + \mu_k$$

③

c) Get $k_1, \ldots, k_n$ iid samples of k

- want to infer temp
- Find $\hat{\Theta}_n$
- Use max liklihood of θ given $k_1, \ldots, k_n$

~~tmp~~

Likelihood function $p_{k_1, \ldots, k_n}(k_1, \ldots, k_n; \theta) = \frac{1}{(Z(\theta))^n} \prod_{i=1}^{n} e^{-k_i/\theta}$

$\uparrow$

by indp.

- lets us take log
- so easier to take deriv
- set $=0$ to maximize

$$\max_{\theta} \text{log liklihood} = -n \log(Z(\theta)) - \frac{1}{\theta} \sum_{i=1}^{n} k_i$$

take deriv of both sides in p.

$$0 = \frac{d}{d\theta} \text{log liklihood} = -n \underbrace{\frac{e^{-1/\theta}}{\theta^2(1-e^{-1/\theta})}}_{\text{is } \frac{d}{d\theta}\left(\log\left(\overbrace{\frac{1}{1-e^{-1/\theta}}}^{\text{is } Z(\theta)}\right)\right)} + \frac{1}{\theta^2} \sum_{i=1}^{n} k_i$$

④

$$\frac{1}{e^{1/\theta}-1} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

Solve for θ to get ML
- non linear
So can't solve eaiser in closed form
trick: linearize it

$$e^{1/\theta} \approx 1 + \frac{1}{\theta} + \underbrace{O(\theta^2 + \ldots)}_{\text{does not matter}}$$

$$\hat{\theta}_n \underset{\uparrow}{\approx} \frac{1}{n}\sum_{i=1}^{n} k_i = \hat{k}_n$$

approx = to ↑ by linearization

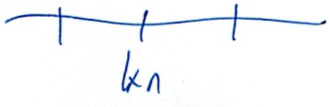
- models units don't agree
- but said 1 implies the other

Can do CLT, var, CI, etc now — dealing w/ sample mean
↩ now can answer this

d) What is # of samples needed so that

$$\frac{\sigma_{\hat{k}_n}}{\hat{k}_n} \text{ is } \leq .01$$

← like a $\frac{\text{noise}}{\text{signal}}$ ratio
$\leq 1\%$



$$= \frac{\sigma_k/\sqrt{n}}{\mu_k} \le .01$$

$$\sqrt{n} \ge \frac{\sqrt{\mu_k^2+\mu_k}}{.01 \cdot \mu_k}$$

$$= 100\sqrt{1+\frac{1}{\mu_k}}$$

If temp is high, than # photons emitted is high

For $\theta >> 1$, $\mu_k >> 1$

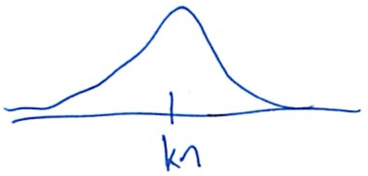
So $\frac{1}{\mu_k}$ goes away

And n should be $\ge 10{,}000$

---

e) Find a 95% CI for $\hat{k}_n$

For $n = 10{,}000$

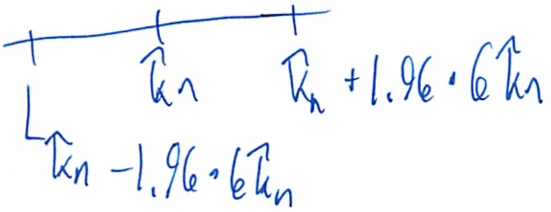


← approx normal

Want upper + lower limit so .95 of values inside

See how many st. dev to left + right from normal table

ⓔ



$$\left[\hat{\lambda}_n - 1.96 \cdot \underset{\overset{\uparrow}{\sqrt{n}}}{.0[illegible]1}\,\hat{\lambda}_n,\ \hat{\lambda}_n + 1.96 \cdot \underset{\overset{\uparrow}{\sqrt{n}}}{.0[illegible]1}\,\hat{\lambda}_n\right]$$

- all very standard
- expect on exam
- will be ~~a~~at least 1 qu
- ~~the~~ mid term comprehensive, but focus on new material

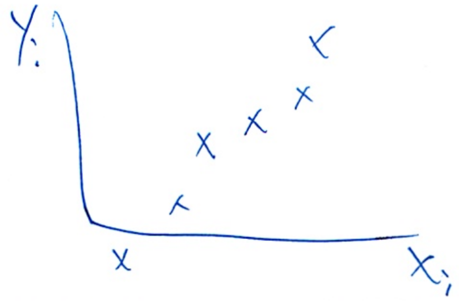
---

2. Regression

- at 1st glance no relation to prob.
- but has statistical interpretations → ML

First do it non-stat way

Construct a model by relating X and Y
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\uparrow\qquad\quad\uparrow$
$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad$ input $\quad$ output



⑦

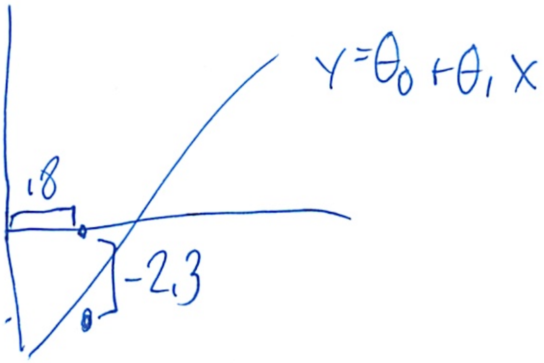
Think x = time
y = position along 1 axis

Want speed/velocity of car

$$\underset{\overset{\uparrow}{\text{pos}}}{y} = \underset{\overset{\uparrow}{\text{inital pos}}}{\theta_0} + \underset{\overset{\uparrow}{\text{velocity}}}{\theta_1} \underset{\overset{\uparrow}{\text{time}}}{x}$$

Given data points on a table

x	.8	...
y	2.3	...



Want vertical descipencies to be as small as possible
Minimize

$$\min \sum_{i=1}^{n} (\overbrace{y_i - \theta_0 - \theta_1 x_i}^{\text{that vertical discrepency}})^2$$

↑simple linear regression model

⑧

take deriv
Set = to 0

$$\frac{\partial}{\partial \theta_0}\sum(\ \ )^2 = 0 = \sum_{i=1}^{n}(y_i-\theta_0-\theta_1 x_i)$$

$$\frac{\partial}{\partial \theta_1}\sum(\ )^2 = 0 = \sum_{i=1}^{n}x_i(y_i-\theta_0-\theta_1 x_i)$$

Now system of 2 eq w/ 2 unknowns

If know θ_0 - plug in, ...
↑ initial pos

If know θ_1, convert to θ_0 w/

$$\hat{\theta}_0 = \bar{Y} - \hat{\theta}_1 \bar{X}$$

↳ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ = output sample mean

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ = input sample mean

Get

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{x})(Y_i-\bar{y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

Divide top + bottom by n

Like a Sample cov

$\approx \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$ ← that is where this ratio comes in
- Coefficent for x - in other things we've seen

For 5 data points

$\bar{X} = 4.94$

$\bar{Y} = 134.38$

$\hat{\theta}_1 = 40.53$
$\hat{\theta}_0 = -65.86$ } coefficents for the straight line

Represents the graph on pg 2 recitation handout

$$Y = 40.53X - 65.86$$

Now try to connect w/ stats

- ~~Construct~~ use ML
- to do dist, CI, etc

<u>ML Interpertation</u>

$$Y_i = \theta_0 + \theta_1 \underset{\substack{\uparrow \\ \underline{\text{not}}\ \text{RV}}}{X_i} + \underset{\substack{\uparrow \text{noise} \\ \sim \text{normal}(0, \sigma^2)}}{W_i}$$

$$\underset{\substack{\uparrow \\ \text{mean}}}{0} \quad \underset{\substack{\uparrow \\ \text{var}}}{\sigma^2}$$

Y_i is a sample of Υ_i

⑩

Likelihood function of $y_1, \dots, y_n$

$$= \text{Constant} \cdot e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 X_i)^2}$$

Maximize liklihood of θ_0, θ_1

= ~~Minimizing~~ max the exponant

= minimizing the expression $\sum_{i=1}^{n} (Y_i - \theta_0 - \theta_1 X_i)^2$
θ_0, θ_1

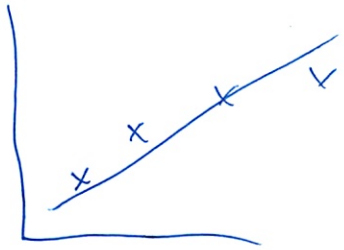
What is the advantage of doing it the states way?

$\hat{\theta}_0, \hat{\theta}_1$ become RVs which get a distribution

But also their var can be estimated

↳ to form CIs

Many other things about regression



Sometimes better w/ higher order polynomial

Instead of assuming linear

- assume quadratic

$Y = \theta_0 + \theta_1 X^2$

Get a better fit

Same formulas

- but w/ $(X_i)^2$ not (x_i)
- so just ^{2} where ever there is an X_i

Or also non 0 slope at y axis

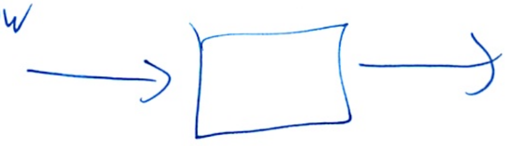
$$Y = \theta_0 + \theta_1 X + \theta_2 X^2$$

nonlinear regression

- can do, but not w/ our formulas
- more complex
- must differentiate 3 parts
- Solve 3 eq, 3 unknowns
- Computer can do it

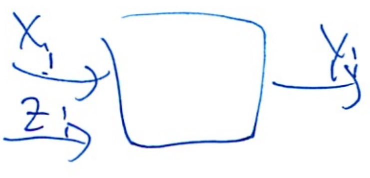
Called multi-parameter regression

Now



Multiple regression

But can also have



or

or even

$Y_i = \theta_0 + \theta_1 X_1 + \theta_{2} Y_i$

12/8

LECTURE 25
Outline

- **Reference:** Section 9.4 Last lecture
- Course VI Underground Guide Evaluations
  `https://sixweb.mit.edu/student/evaluate/6.041-f2010`
  `https://sixweb.mit.edu/student/evaluate/6.431-f2010`

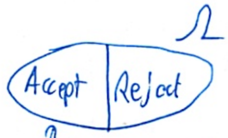
more classical stats

- Review of simple binary hypothesis tests test hypothesis
  - examples
- Testing composite hypotheses
  - is my coin fair?
  - is my die fair?
  - goodness of fit tests

Simple binary hypothesis testing

No prior beliefs about hyp

- **null hypothesis** H_0:
  $X \sim p_X(x; H_0)$ [or $f_X(x; H_0)$] - Vector of several RVs
- **alternative hypothesis** H_1:
  $X \sim p_X(x; H_1)$ [or $f_X(x; H_1)$]
- Choose a **rejection region** R;
  reject H_0 iff data $\in R$



depends where data falls

- Likelihood ratio test: reject H_0 if

$$\frac{p_X(x; H_1)}{p_X(x; H_0)} > \xi \quad \text{or} \quad \frac{f_X(x; H_1)}{f_X(x; H_0)} > \xi$$

compare likelihood ratio to some ~~constant~~ threshhold param

  - fix false rejection probability α; (e.g., $\alpha = 0.05$)
  - choose ξ so that P(reject H_0; H_0) $= \alpha$

- but how chose it?
- Set a ~~parameter~~ criteria
- then solve to get ξ

; How design acception region - have a choice

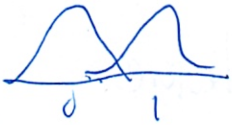
Two types of errors
- false negitives
- false positive

Example (test for normal mean)

- n data points, i.i.d.
  H_0: $X_i \sim N(0,1)$
  H_1: $X_i \sim N(1,1)$

two hypothses

both normal but different means



- Likelihood ratio test; rejection region:

$$\frac{(1/\sqrt{2\pi})^n \exp\{-\sum_i (X_i-1)^2/2\}}{(1/\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/2\}} > \xi$$

↑ look at likelihood ratio

Write down joint densities under $\frac{H_1 \text{ top}}{H_0 \text{ bottom}}$ - evaluated at observed values x

  - algebra: reject H_0 if: $\sum_i X_i > \xi'$

- Find ξ' such that

$$\mathbf{P}\Big(\sum_{i=1}^{n} X_i > \xi'; H_0\Big) = \alpha$$

  - use normal tables

Compare what we got w/ threshhold

$$-\sum_i (x_i - 1)^2 + \sum x_i^2 > \log \xi$$

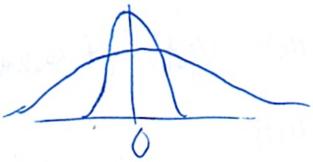
$$2\sum x_i - n > \log \xi$$

If sum of X_i s big - than choose H_1 - ie reject H_0

Example (test for normal variance)

- n data points, i.i.d.
  H_0: $X_i \sim N(0,1)$
  H_1: $X_i \sim N(0,4)$

here same mean



Simple since both hyp are fully specified probabilistic models

- Likelihood ratio test; rejection region:

$$\frac{(1/2\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/(2\cdot 4)\}}{(1/\sqrt{2\pi})^n \exp\{-\sum_i X_i^2/2\}} > \xi$$

← same cookbook procedure

  - algebra: reject H_0 if $\sum_i X_i^2 > \xi'$

} take log, like before

- Find ξ' such that

$$\mathbf{P}\Big(\sum_{i=1}^{n} X_i^2 > \xi'; H_0\Big) = \alpha$$

prob of rejecting hypothsis

  - the distribution of $\sum_i X_i^2$ is known (derived distribution problem)
  - "chi-square" distribution; ← form not all that important
    tables are available

look up 95%th tile to find threshhold

Composite hypotheses

- Got $S = 472$ heads in $n = 1000$ tosses; is the coin fair?
  - $H_0 : p = 1/2$ versus $H_1 : p \neq 1/2$

    ↑ coin fair
- Pick a **"statistic"** (e.g., S)
- Pick shape of **rejection region** (e.g., $|S - n/2| > \xi$)

  reject if very different
- Choose **significance level** (e.g., $\alpha = 0.05$)

  ↓ threshold
- Pick **critical value** ξ so that:

  $$\mathbf{P}(\text{reject } H_0; H_0) = \alpha$$

  Using the CLT: ↑ H_0 is fully specified

  $$\mathbf{P}(|S - 500| \le 31; H_0) \approx 0.95; \quad \xi = 31$$

  look at normal tables, say S is normal
- In our example: $|S - 500| = 28 < \xi$

  H_0 **not rejected** (at the 5% level)

  ↑ delebirate choice of words

Can't say H_1 is accepted

And will not be exactly .500 often

But are the obs compatable (close) to H_0

So H_0 could be true (not rejected)

H_0 is fully specified

← not fully specified model H_1

lots of alterntative models

- $p = .6$
- $p = .7$
- $p = .499$

} each is fully specified

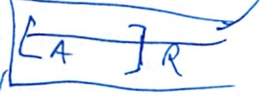
If

$H_0 = p = 1/2$

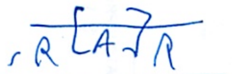
$H_1 = p > 1/2$

then rejection region

if $S \geq \frac{n}{2} + \frac{?}{?}$



current region



Can't write likelihood ratio anymore

- Since many canidate distributions

Build test on statistic

- Function of data X_is
- Compress into 1 #
- here $S = 472$ # of heads
- not always clear what to use

- Broader philosophical: Theories can't be proven correct - only that has not been rejected

Is my die fair?

- Hypothesis H_0:

  $\mathbf{P}(X = i) = p_i = 1/6, \ i = 1, \ldots, 6$
- Observed occurrences of i: N_i

  how many times each was observed

  Here what should S be?
- Choose form of rejection region; chi-square test:

  ✓ value i shows up $\frac{n}{6}$ times supposedly

  - what actually happens?

  $$\text{reject } H_0 \text{ if } T = \sum_i \frac{(N_i - np_i)^2}{np_i} > \xi$$

  ↑ take squares

  divide by factor - different weights depending on P_i

  - but does not matter here - constant
- Choose ξ so that:

  $$\mathbf{P}(\text{reject } H_0; H_0) = 0.05$$

  $$\mathbf{P}(T > \xi; H_0) = 0.05$$

  look backwards

  - if big values are far away

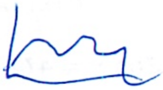
  choose like before
- Need the distribution of T: (CLT + derived distribution problem)
  - for large n, T has approximately a chi-square distribution
  - available in tables

↑ formula or table

- if P_is equal - what type of dist is T?
- N_i is binomial (1/6)
- CLT is approx same as normal - 0 mean of normal
- take square of a normal - sum is chi squared (earlier today)

More complicated versions of this

Hyp: RVs distributed by this PDF



Do I have the correct pdf?

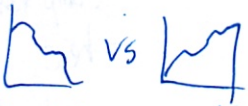
- Partition the range into bins
  - np_i: expected incidence of bin i (from the pdf)
  - N_i: observed incidence of bin i
  - Use chi-square test (as in die problem)

- Split into bins
- PDF ~~[illegible]~~ predicts prob for each bin
- Count how often got a result in ~~some~~ each bin
- compare w/ expected # of observences
- add over i's

$$\sum_i \frac{(N_i - np_i)^2}{np_i}$$

that RV is chi squared

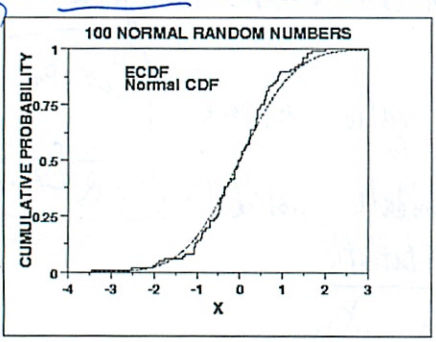
- can't dinguish b/w same PDF for a certain region



- Only tells you if compatable w/ hyp — survied so far
- not if true

- Kolmogorov-Smirnov test: form **empirical CDF**, $\widehat{F}_X$, from data

Other more sophisticated methods
- If want no binning
- compare CDF
- what fraction of data below some #



(http://www.itl.nist.gov/div898/handbook/)

- $D_n = \max_x |F_X(x) - \widehat{F}_X(x)|$
- $P(\sqrt{n}D_n \geq 1.36) \approx 0.05$ ✓ prob false rejection

- much approch prob of falling below that #
- need to find threshhold value
- once have it, know what test must be
  - where curves far enough apart

What else is there?

In statistics

want informative 1D summary
want to be able to analize
- ie form tables

- Systematic methods for coming up with shape of rejection regions likelihood ratio test
- Methods to estimate an unknown PDF (e.g., form a histogram and "smooth" it out) where coming from?
- Efficient and recursive signal processing
  can you do it quickly?
- Methods to select between less or more complex models
  - (e.g., identify relevant "explanatory variables" in regression models)
- Methods tailored to high-dimensional unknown parameter vectors and huge number of data points (data mining)
- etc. etc....

- no unique correct ans - room for judgement - art
if get results you want - you stick w/ it
- he hopes the better methods will survive
- also need it to be computationally implementable

~~[illegible]~~ ~~[illegible]~~
- can always find very complex model that fits data exactly
- but that does not match reality

↑ where work is today
takes many parameters to explain data point
huge # of data points
- advance class next semester

Chap 9 Classical Statistical Inference

12/6

Previous chap: Bayesian approach to inference

-Unknown params modeled as RVs

but here θ is known (not random)

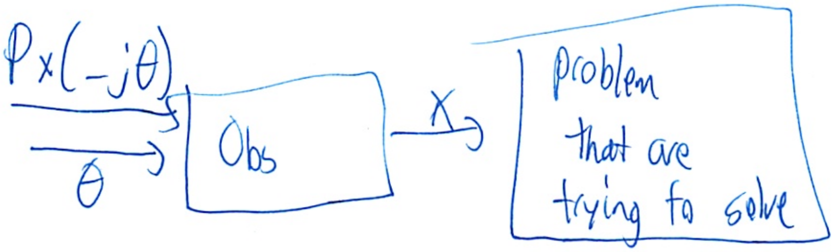
Obs X is random

$p_X(x;\theta)$ or $f_X(x;\theta)$ depends on value of θ

↳ multiple canidate models

One for each value of θ

So want to find best (out of all canidate models)



$E_\theta[h(X)]$ = expected value of RV $h(X)$ as function of θ

$P_\theta(A)$ = prob of event A

↳ functional dependence - not conditioning

Types of problems

- parameter estimation - estimates that are nearly correct under any possible value of unknown param.
- hypothesis testing - unknown param takes finite # of values m $(m \geq 2)$ - want to pick one w/ least error

②

- Significance testing - want to accept/reject 1 hypothsis while keeping prob of false rejection small

Inference Methods

- Maximum Liklihood (ML) - parameter that makes observed data most likly - max. chance of obtaining it
- Linear Regression - linear relation that minimizes sum of square of error b/w line + data
- Likelihood Ratio Test - Given two hyp, compare their relative (ratio) of liklyhood
- Significance testing - Given a hyp, reject if obs data falls outside of certain rejection region

equivilant

MAP

9.1 Classical Parameter Estimation

θ is unknown constant

ML is classical equivilant of MAP

$X = X_1, \ldots, X_n$ = observations

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

distribution of X depends on θ, also dist of Θ depends of θ

$\hat{\theta}$ = estimate (the actual value)

③

Some times interested in ~~the~~ when $n \to \infty$, $n = \#$ of obs

$\hat{\Theta}_n$ = estimator

↳ actually seq of estimators – one for each n

↳ mean = $E_\theta(\hat{\Theta}_n)$ $\Big)$ # functions of θ

↳ var = $var_\theta(\hat{\Theta}_n)$

$\tilde{\Theta}_n$ = estimation error = $\hat{\Theta}_n - \theta$

bias = $b_\theta(\hat{\Theta}_n)$ = expected value of estimation error

$= E_\theta[\hat{\Theta}_n] - \theta$

↳ E[bias] depend on θ

↳ var(bias) " " "

↳ estimation error also depends on observences $X_1, X_2, \ldots, X_n$

$\hat{\Theta}_n$ is unbiased if $E_\theta[\hat{\Theta}_n] = \theta$ for every possible θ

$\hat{\Theta}_n$ is asymptotically unbiased if $\lim_{n \to \infty} E_\theta[\hat{\Theta}_n] = \theta$ for every possible θ

$\hat{\Theta}_n$ is consistent if seq $\hat{\Theta}_n$ converges to true value of θ in prob for every possible value of θ

④

Estimator will not be = to θ exactly

↳ So estimation error will be non 0

But if estimation error avg is 0 → unbiased

If only as $n\uparrow$, then asymptotically unbiased

Also interested in size of estimation error

↳ mean squared error $E_\theta[\Theta_n^2]$

$$E_\theta[\Theta_n^2] = b_\theta^2(\hat{\Theta}_n) + var_\theta(\hat{\Theta}_n)$$

In many problems tradeoff b/w the two terms on right hand side

- ie if ↓ var ~~bar~~ bias ↑

(can I see a graphical example of that?)

- Goal is ~~a~~ to keep both small

ML Estimator

$X = X_1, \ldots, X_n$ = vector of obs

describe w/ joint PMF $p_X(X;\theta)$

- form depends on an unknown (scalar or vector) θ

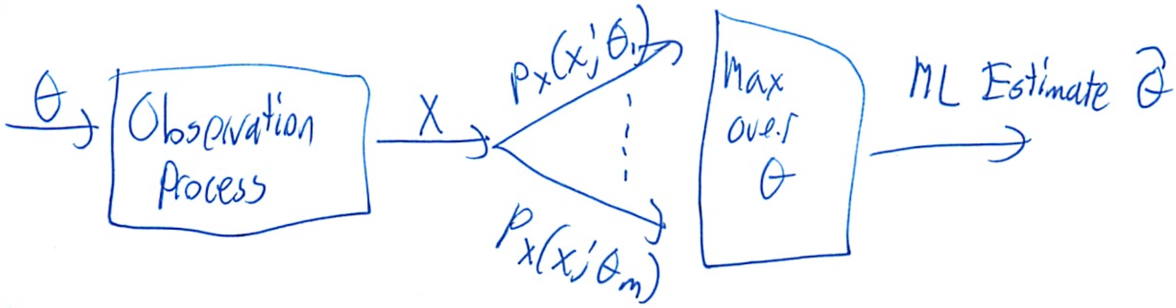
Observe a certain value $x = (x_1, \ldots, x_n)$

Maximize $p_X(x_1, \ldots, x_n; \theta)$ over all θ

$$\hat{\Theta}_n = \arg\max_\theta p_X(x_1, \ldots, x_n; \theta)$$

(or f if continuous

⑤



In many cases observations X_i are assumed to be ind
So likelihood function of the form

$$p_X(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} p_{X_i}(x_i; \theta)$$

But eaiser to max log → log-likelihood function over θ

$$\log p_X(x_1, \dots, x_n; \theta) = \log \prod_{i=1}^{n} p_{X_i}(x_i; \theta)$$

$$= \sum_{i=1}^{n} \log p_{X_i}(x_i; \theta)$$

likelihood → is not parameter that unknown param = θ
↳ instead prob that observed value x can arise when parameter is = to θ
"What is the value of θ under which the obs we have seen are most likely to arise"

with a flat prior they are the same

if one-to-one function - just apply function to estimate

(skipping examples)

⑥ 12/7

Estimation of Mean and Var of RV

- simple, but important problem of estimating mean + var
- we don't need to know about distribution

- Sample mean – most natural estimator of θ

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

  - unbiased since $E_\theta[M_n] = E_\theta[X] = \theta$
  - MSE $= \text{var} = \frac{v}{n} \leftarrow$ common var of X_i
    - does not depend on θ
    - by WLLN → converges to θ in probability = consistent
  - Sample mean not nessarily estimator w/ smallest var

- two var estimators

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

- coincides w/ ML if X_is are normal
- biased, but asymptotically unbiased

- unbiased

for large n the two estimators coincide

(?but which to use?)

- one is based on ML
- second is scalled to be unbiased

Confidence Interval

Have estimator $\hat{\Theta}_n$ of unknown param θ

So confidence θ is within interval

Confidence level $1-\alpha$

Set bands

↙ confidence interval

$$P_\theta\left(\hat{\Theta}_n^- \le \theta \le \hat{\Theta}_n^+\right) \ge 1-\alpha$$

(I think I get it conceptually - just need practice)

for every possible value of θ

↑ (this kinda confuses me how they worded it)

remember in classical stats its the ~~cons~~ interval that is random

θ is fixed

Well 95% of intervals will include θ

Usually constructed by forming an interval around an estimator - want smallest possible width - but this depends on θ - but this is usually asymptotically normal and asymptotically unbiased

So $$\frac{\hat{\Theta}_n - \theta}{\sqrt{\mathrm{Var}_\theta(\hat{\Theta}_n)}}$$

CDF ~~of normal~~ approaches st normal as $n \uparrow$ for every θ

Estimator Var Approx

- if var known, just use that
- if not known, need to estimate
- Using the unbiased estimator

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

- estimate var $\frac{v}{n}$ of sample mean by $\frac{S_n^2}{n}$
- So for a given α, use these estimates + CLT to conduct approx $1-\alpha$ confidence interval
- So $\left[\hat{\Theta}_n - z\frac{\hat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z\frac{\hat{S}_n}{\sqrt{n}}\right]$

where z is obtained from

$$\Phi(z) = 1 - \frac{\alpha}{2}$$

and normal tables

$$\Phi(1.96) = .975 = \frac{1-\alpha}{2}$$

↪ backwards on table

- So get

$$\left[\hat{\Theta}_n - 1.96\frac{\hat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + 1.96\frac{\hat{S}_n}{\sqrt{n}}\right]$$

- two diff approx in effect
  - treating $\hat{\Theta}_n$ as if normal
  - replacing var w/ estimate
- allways only an approx CI
  - $\hat{S}_n^2$ only approx to true var v
- The RV $T_n = \dfrac{\sqrt{n}(\hat{\Theta}_n - \theta)}{\hat{S}_n}$ is not normal
  - but PDF of T_n does not depend on θ or v
  - So can compute explicitly
  - t-distibution w/ n−1 degrees of freedom
  - symettric + bell shaped - but more spread out w/ heavier tails
- When X_i normal, n relativly small

$$\left[\hat{\Theta}_n - z\frac{\hat{S}_n}{\sqrt{n}}, \hat{\Theta}_n + z\frac{\hat{S}_n}{\sqrt{n}}\right]$$

- z obtained from relation

$$\Psi_{n-1}(z) = 1 - \frac{\alpha}{2}$$

So when n is large ($n \geq 50$) t-distribution is close to normal → so can use normal tables
(did not do in class I think)

Otherwise are t tables → CDF $\Psi_{n-1}(z)$
w/ # of degrees of freedom
desired tail prob
and z value

Diff estimators for var possible

- if Bernouli $V = \theta(1-\theta)$ of X
  So var $= \hat{\theta}_n(1-\hat{\theta}_n)$
- as n ↑ to ∞ $\hat{\theta}_n \to \theta$ in prob
  So $\hat{\theta}_n(1-\hat{\theta}_n) \to V$
- Or say that $\theta(1-\theta) \leq 1/4$ for $\theta \in [0,1]$ and use $1/4$ as conservative var estimate

9.2 Linear Regressions

- building model of relation b/w 2 or more variables of interest
- can explain it simply
- or under guise of probability

(did today in recitation)

Will start off w/ two variables

- have X = years of edu
  Y = income
- have data pairs (x_i, y_i)
  $i = 1, \ldots, n$

x_i = years of edu, y_i = income of <u>ith</u> person

~~Rep~~ Represent' linearly

$$Y \approx \underset{\uparrow}{\theta_0} + \underset{\uparrow}{\theta_1} x$$

unknown params to be estimated

Given estimates $\hat{\theta}_0, \hat{\theta}_1$ of resulting params, y_i corresponds to x_i as predicted by the model

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$$

The "error" y_i is given by

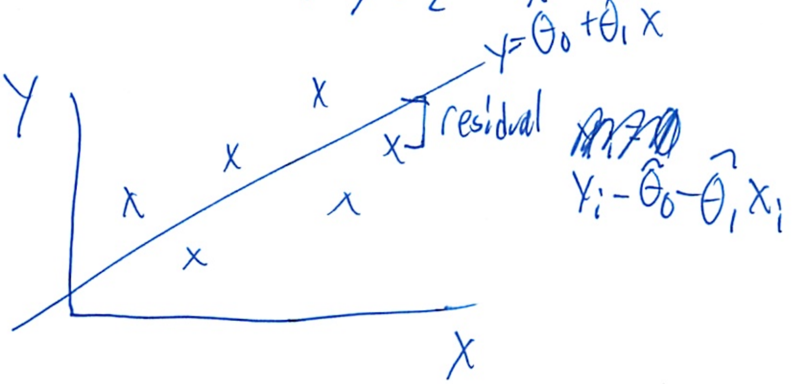
$$\tilde{y}_i = y_i - \hat{y}_i \qquad \text{called the } i\text{th } \underline{\text{residual}}$$

Want to choose estimates so have a small residual

So minimize the square of the residuals

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Over all θ_1 , θ_2



Postulated linear model may or may not be true

- if relation is actually nonlinear
- in order to use this model we assumed linearity

To derive and get $\hat{\theta}_0$ and $\hat{\theta}_1$ - take partial derivs + set = 0

$$\theta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

⑬

Can be justifed many ways
(skipping)

Bayesian Linear Regression

12/8

Can explain linear Regression w/ prob Basian framework

$x_1, \dots, x_n$ = given #

$y_1, \dots, y_n$ = Observed values of Vector $Y = (Y_1, \dots, Y_n)$
of RVs that obey linear relation

$Y_i = \Theta_0 + \Theta_1 x_i + W_i$

↑ $\Theta = (\Theta_0, \Theta_1)$ is parameter to be estimated

↑ $W_1, \dots, W_n$ are iid RVs
w/ mean 0 and known var σ^2

$\Theta_0, \Theta_1, W_1, \dots, W_n$ are indp.

Θ_0, Θ_1 have mean 0 and var σ_0^2, σ_1^2

Derive a Bayesian estimator based on MAP

Assume $\Theta_0, \Theta_1, W_1, \dots, W_n$ are normal RV

Maximize θ_0, θ_1 over posterior PDF

$$f_\Theta(\theta_0, \theta_1) f_{Y|\Theta}(y_1, \dots, y_n \mid \theta_0, \theta_1)$$

⑭

Divided by a positive normalization constant that does not depend on (θ_0, θ_1)

This is

$$C \cdot \exp\left\{-\frac{\theta_0^2}{2\sigma_0^2}\right\} \cdot \exp\left\{-\frac{\theta_1^2}{2\sigma_1^2}\right\} \cdot \prod_{i=1}^{n} \exp\left\{-\frac{(y_i - \theta_0 - x_i\theta_1)^2}{2\sigma^2}\right\}$$

where C = normalizing constant

Equivalently, ~~normalize over~~ minimize over θ_0, θ_1

$$\frac{\theta_0^2}{2\sigma_0^2} + \frac{\theta_1^2}{2\sigma_1^2} + \sum_{i=1}^{n} \frac{(y_i - \theta_0 - x_i\theta_1)^2}{2\sigma^2}$$

↑ Same as w/ classical

↑ would be identical if σ_0, σ_1 so large, these terms disappear

Take partial derivs

Set $= 0$

$$\hat{\theta}_1 = \frac{\sigma_1^2}{\sigma^2 + \sigma_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2} \cdot \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\theta}_0 = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} (\bar{y} - \hat{\theta}_1 \bar{x})$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Some remarks - skipping

Multiple Linear Regression

- So far have had single explanatory variable x
  - ↳ simple regression
- but often multiple underlying or explanitory variables
  - ↳ multiple regression

if had triplets of data (x_i, y_i, z_i) and wanted θ_i

$$y \approx \theta_0 + \theta_1 x + \theta_2 z$$

then minimize

$$\sum_{i=1}^{n} (y_i - \underset{\uparrow}{\theta_0} - \underset{\uparrow}{\theta_1} x_i - \underset{\uparrow}{\theta_2} z_i)^2$$

minimize over

no limit on possible explanitory variables

Non Linear Regression

- if we assume model is nonlinear

$$y \approx \underset{\uparrow \text{ given function}}{h}(x; \underbrace{\theta}_{\text{param to be estimate}})$$

- given data pairs (x_i, y_i) $i = 1, \ldots, n$
- Seek value of θ that minimizes sum of squared residuals

$$\sum_{i=1}^{n} (y_i - h(x_i; \theta))^2$$

But no general closed form solution

But can try w/ enough computational power

- can do via ML estimation

$$Y_i = h(x_i; \theta) + W_i$$

↑ iid normal w/ 0 mean

- liklihood function takes form

$$f_Y(y;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(y_i - h(x_i;\theta))^2}{2\sigma^2}\right\}$$

- Heteroskedasticity
  - var of W_i could vary w/ x_i
  - So noise at the end overwhelms noise at the begining
  - Can adjust weighting to avoid
- Nonlinearity
  - ~~ph~~ Sometimes pick a non linear model
- Multicollinearity
  - if x, z closly related model may not be able to distinguish b/w the two
- Overfitting
  - fits data well, but model useful
  - if your polynomial has too high of an order
  - must be 5-10 x data points than paramter being estimated

Causality

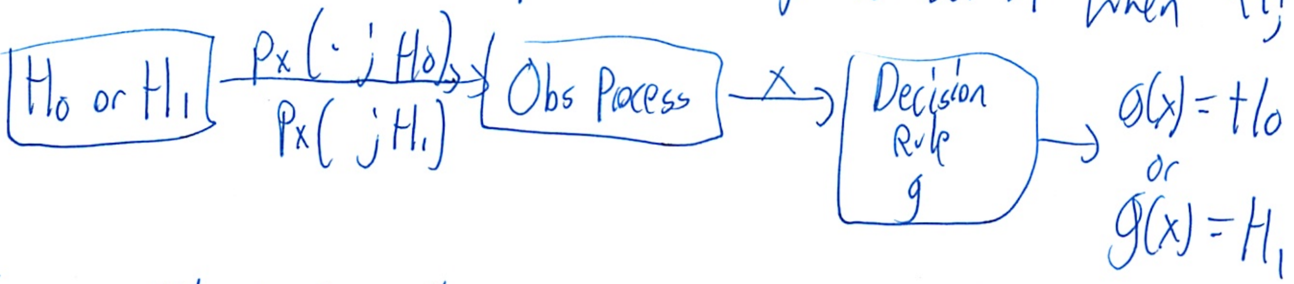
the holy grail of science - causation vs correlation

⑰

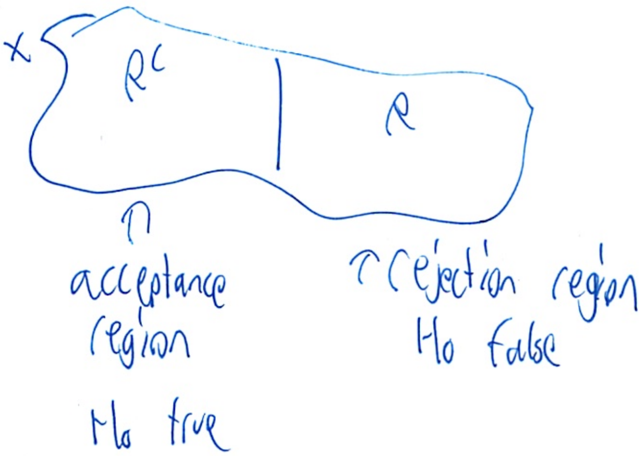
9.3 Binary Hypothesis Testing

- back to choosing b/w 2 hyp
- but unlike 8.2 section - no prior prob
- like an infrence problem where θ takes 2 values
  - but call H_0 = null / default ← prove or disprove
    H_1 = alternate
- Observation is vector $X = (X_1, ..., X_n)$ of RV
  - dist depends on hyp

$P(X \in A; H_j)$ to denote prob X belongs to set A when H_j is true



- Can partition observations into 2 subsets



⑱

Two possible types of errors

a) False positive/Type I/false rejection

Reject H_0, even though H_0 is true

Prob of happening

$$\alpha(R) = P(X \in R; H_0)$$

b) False negitive/ Type II/ False acceptance

Accept H_0 even though H_0 is false

$$\beta(R) = P(X \notin R; H_1)$$

To decide shape of region, minimize prob of error w/ MAP rule!

Given observed value x of X, declare $\Theta = \theta_1$ to be true if

$$p_\Theta(\theta_0)\, p_{X|\Theta}(x|\theta_0) < p_\Theta(\theta_1)\, p_{X|\Theta}(x|\theta_1)$$

Rewrite as Liklihood Ratio $L(x)$

$$L(x) = \frac{p_{X|\Theta}(x|\theta_1)}{p_{X|\Theta}(x|\theta_0)}$$

Declare $\Theta = \theta_1$ to be true if realized value x of obs vector X satisfies

$$L(x) > \xi$$

Where ξ is the critical value

$$\xi = \frac{P_\Theta(\theta_0)}{P_\Theta(\theta_1)}$$

If X is contineous, same but w/ f

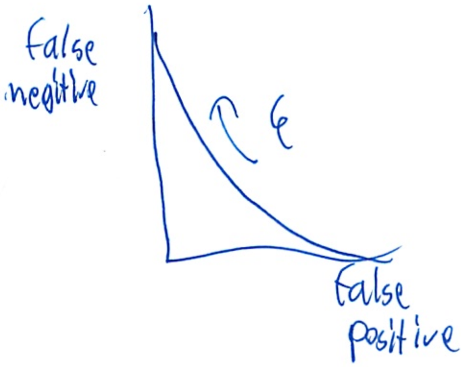
So consider rejection regions of the form

$$R = \{x \mid L(x) > \xi\}$$

ξ chosen through other considerations

$\xi = 1$ is the ML rule

ξ is a tradeoff b/w both types of error



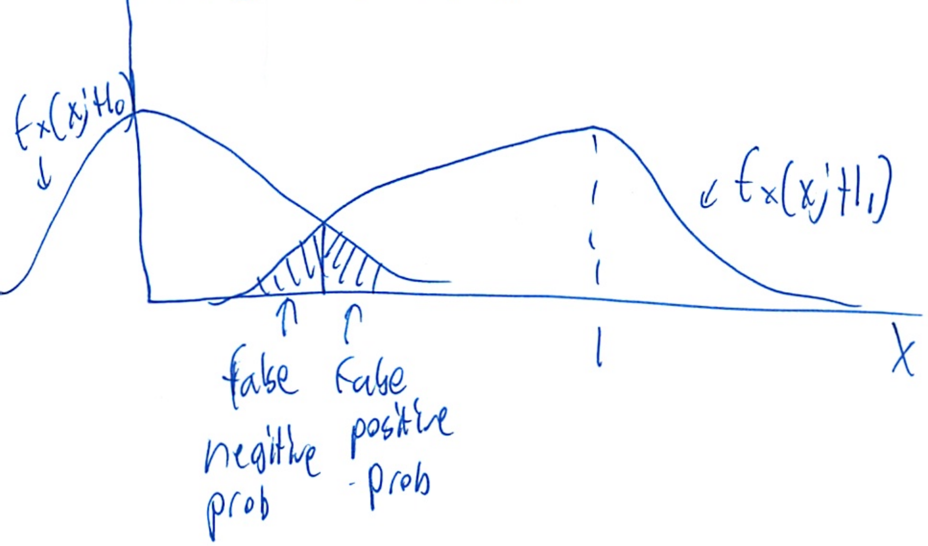
So pick it using the Likelihood Ratio Test (LRT)

- start w/ target α for false positive prob
- Pick ξ so false ~~prob~~ pos prob = to α

$$P(L(x) > \xi; H_0) = \alpha$$

- once x observed, reject H_0 if $L(x) > \xi$

(skipping requirements)



(I don't really get this chart)

(¿If it is Ho will fall - but if in false pos teritory will be said to be H1, when it is Ho

if falls in normal Ho or false negitive prob teritory say is ok)

(kinda cool how they represent that)

Neyman-Pearson Lemma

Consider a particular choice of ξ in the LRT which results in

$P(L(x) > \xi; H_0) = \alpha$ $P(L(x) \le \xi; H_1) = \beta$

Suppose that some other test w/ rejection region R achives smaller or = false positive prob

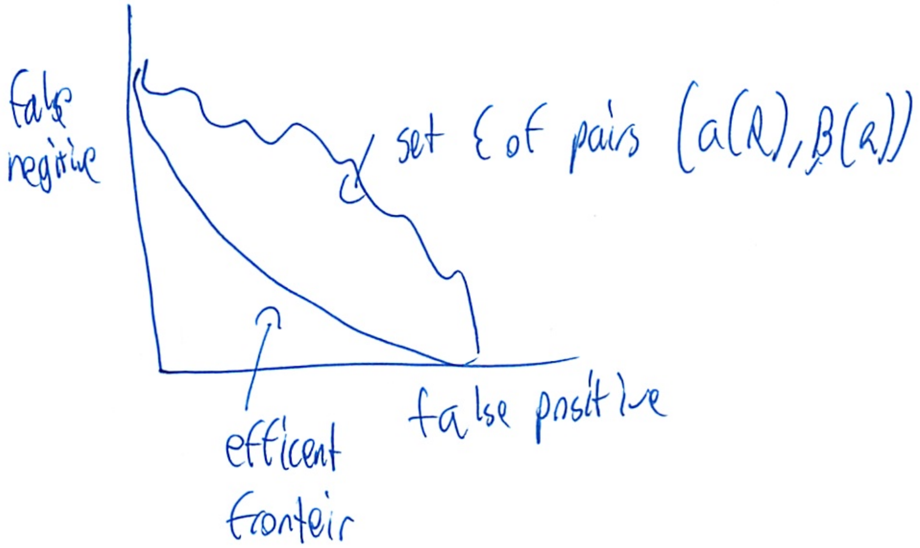
$P(X \in R; H_0) \le \alpha$

Then $P(X \notin R; H_1) \ge \beta$

w/ strict inequality $P(X \notin R; H_1) > \beta$ when $P(X \in R; H_0) < \alpha$

㉑

(I don't get that at all!)



Lemma states all pairs $(a(\xi), \beta(\xi))$ lie on efficent fronteir

(still don't get - don't think we got into in class)

9.4 Significance Testing

Hypothesis testing problems do not always have 2 well specified alternatives, so can't use 9.3

This is more general

A lot more "art"/judgement here

(more complex questions)

default hyp = H_0 = null hypothesis

Want to determine based on observations $X = (X_1, ..., X_n)$ whether hyp should be accepted or rejected

Will restrict discussion to models w/ following characistics

a) Parametric models – obs have dist governed by joint PMF/PDF
Completly determined by Θ, belongs to M set of parameters

b) Simple Null hypothesis – ~~true~~ asserts true value of Θ is = to a given element Θ_0 of M

c) Alternate hypothesis – H_1 that H_0 is not true

(went over this in lecture today)

Coin toss example

coin tossed $n = 1000$ times

θ = unknown probability of each toss

set of all possible params $M = [0, 1]$

H_0 = null hyp = the coin is fair $\theta = 1/2$

alt hyp = $\theta \neq \frac{1}{2}$

Observe $X_1, \ldots, X_n$ tosses

↳ $X_i = 0$ or 1 heads

$S = X_1 + \ldots + X_n$

Use decision rule

reject H_0 if $\left|S - \frac{n}{2}\right| > \xi$

critical value to be determined

We have defined R (Rejection region)

↳ set of data vectors that lead to rejection of null hyp

Choose ϵ so prob false positive / false rejection $= \alpha$

$$P(\text{reject } H_0;\ H_0) = \alpha$$

↑ <u>significance level</u>, here $\alpha = .05$

Now need to make some choices. Some prob calculations needed to determine critical value ϵ

Under null hyp, S is binomial w/ p $n = 1000$, $p = \frac{1}{2}$

Use normal approx to binomial + normal tables

$\epsilon = 31$

If observed $S = s = 472$

$$|s - 500| = |472 - 500| = 28 \le 31$$

and Ho is <u>not rejected at 5% significance level</u>

↑ only say that observed S does not provide strong evidence against hypothesis Ho

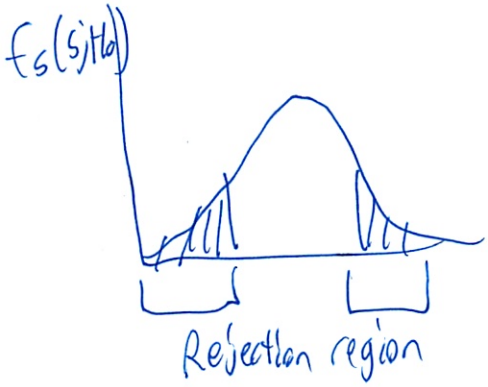
↓

~~Sta~~ Significance Testing Methodology

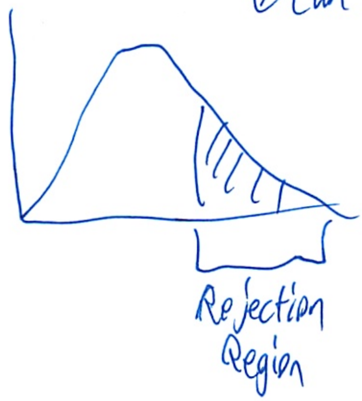
A stat. test of hyp $H_0 = \theta = \theta^*$ is performed based on obs $X_1, \ldots, X_n$

1. Chose a statistic S that is a scalar RV
   - choose some ~~functi~~ function h $R^n \to R$ resulting in statistic $S = h(X_1, \ldots, X_n)$
2. Determine shape of rejection region
   - Specify set of values of S for which H_0 will be ~~rejectd~~ rejected as a function of critical value ξ
3. Choose the significance level
   - ie desired prob ~~α~~ α of a false rejection of H_0
4. Choose the critical value ξ so prob of false positive $\approx \alpha$
   ~~A~~ Rejection region is determined
5. Once $x_1, \ldots, x_n$ are observed
   i) Calculate statistic $s = h(x_1, \ldots, x_n)$
   ii) Reject hyp H_0 if s is in rejection region

1. No right way to Choose S
   - Sometimes obvious
   - sometimes need to generalize likelihood ratios
   - need to make sure it is easy to calculate
2. ~~Set~~ values of S under which H_0 not rejected is usually an interval surrounding peak dist of S under H_0



or



↙ can have 1 sided

3. Pick ~~a~~ α to balance tradeoff
4. Step 4 (previous page) only place prob. calculations used
   - requires dist of $L(x)$ ~~S~~
   - Sometimes $\log L(x)$
   - ~~it~~ often S can not be found in closed form
   - often need approx like CLT
     - but only if n is large
   - may estimate S by simulation

Say H_0 is rejected at the α significance level

Does not mean prob of H_0 being true is less than α

- Instead will have false rejection/False positive a fraction α of the time

---

Staticians often replace step 3+4 (2 pages ago) w/ pvalue

p-vale $= \min\{\alpha \mid H_0 \text{ would be rejected at the } \alpha \text{ sig level}\}$

- at the threshold b/w rejection + non-rejection
- ie null hyp would be rejected at 5% sig level if

p-vale $< .05$

(skipping examples)

Generalized Liklihood Ratio + Goodness of Fit Tests

- test whether a given PMF conforms w/ observed data
  = goodness of fit
- Use it as an introduction to general methodology for significance testing in face of composite alt hypothesis
- Consider RV that takes values in finite set $\{1, \dots, m\}$
- Let θ_k be prob of outcome k
- Dist of RV is described by vector param $\theta = (\theta_1, \dots, \theta_m)$
- Consider hyp $H_0: \theta (\theta_1^*, \dots, \theta_m^*)$ $H_1: \theta \neq (\theta_1^*, \dots, \theta_m^*)$

㉗

θ^*_k are non neg # that sum to 1

Draw n ind. samples of RV

N_k = # of samples that result in outcome k

~~Dist of RV~~

Obs $X = (N_1, \ldots, N_m)$

denote realized value by $x = (n_1, \ldots, n_m)$

$N_1 + \ldots + N_m = n_1 + \ldots + n_m = n$

Do <u>generalized liklihood ratio test</u>

1. estimate a model by ML
   - ie determine a param vector $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_m)$ that maximizes the likilihood fn $p_X(x; \theta)$ over all vectors θ
2. Cary out a LRT that compares liklihood $p_X(x; \theta^*)$ under H° to liklihood $p_X(x; \hat{\theta})$ corresponding to estimated model. Ie form

$$\frac{p_X(x; \hat{\theta})}{p_X(x; \theta^*)}$$

and if it exceeds critical value ξ, reject H₀

Choose ξ so prob false rejection/positive ≈ α

㉘

(we did not do this in class I believe)

Basically we are asking

- is there a model compatable w/ H_1 that provides a better explination for the observed data than that provided by model corresponding to H_0?
- To ans: Compare likelihood under H_0 to largest possible liklihood under models compatable w/ H_1

First step: ML estimation → involves a maximization over the set of prob. distributions $(\theta_0, \ldots \theta_m)$

PMF of obs vector X is multinomial

$$p_X(x;\theta) = c\,\theta_1^{n_1} \ldots \theta_m^{n_m}$$

↑normalizing constant

Eaiser to work w/ Log liklihood

$$\log p_X(x;\theta) = \log c + n_1 \log \theta_1 + \ldots + n_{m-1} \log \theta_{m-1} + n_m \log(1 - \theta_1 - \ldots - \theta_{m-1})$$

Set derivs of $\theta_1, \ldots \theta_{m-1}$ = to 0

$$\frac{n_k}{\hat\theta_k} = \frac{n_m}{1 - \hat\theta_1 - \ldots - \hat\theta_{m-1}} \text{ for } k = 1, \ldots, m-1$$

↑all the ratios must be =

$$\hat\theta_k = \frac{n_k}{n} \quad k = 1, \ldots, m$$

These are correct ML estimates even if some of the $n_k = 0$

- so corresponding $\hat{\theta}_k$ are also 0

Generalized Form

reject H_0 if $$\frac{p_X(x;\hat{\theta})}{p_X(x;\theta^*)} = \prod_{k=1}^{m} \frac{(n_k/n)^{n_k}}{(\theta^*_k)^{n_k}} > \xi$$

Take log to simplify

reject H_0 if $$\sum_{k=1}^{m} n_k \log\left(\frac{n_k}{n\theta^*_k}\right) > \log \xi$$

Need to determine ξ by taking into account α - sig level

$$P(S > \log \xi ; H_0) = \alpha$$

where $$S = \sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n\theta^*_k}\right)$$

But dist of S under H_0 not readily availble - can only simulate

But if n is large → can simplify

- observed freq $\hat{\theta}_k = \frac{n_k}{n}$ will be close to θ^*_k under H_0 w/ high prob

Second order taylor series expansion shows that our statistic S can be app. well by $\frac{T}{2}$, where

$$T = \sum_{k=1}^{m} \frac{(N_k - n\theta^*_k)^2}{n\theta^*_k}$$

When n is large, it is known under Hyp H_0 the dist of T (and dist $2S$) approaches a so-called "χ^2" dist w/ $m-1$ degrees of freedom

- dist avalible in tables
- approx correct values of $P(T > \gamma; H_0)$ or $P(2S > \gamma; H_0)$ can be obtained from χ^2 table
- can use to determine a suitable critical value that corresponds to given significance level α

So have test for large values of n

Chi Square Test

- Use statistic, ~~or related statisti~~

$$S = \sum_{k=1}^{m} N_k \log\left(\frac{N_k}{n\theta^*_k}\right)$$

(or related statistic T)

and rejection region

reject H_0 if $2S > \gamma$

(or $T > \gamma$)

- Critical value γ is determined from CDF tables for χ^2 dist w/ $m-1$ degrees of freedom so that

$$P(2S > \gamma; H_0) = \alpha \quad \text{given stat. level}$$

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

Problem Set 11
Never Due
Covered on Final Exam

1. **Problem 7, page 509 in textbook**
Derive the ML estimator of the parameter of a Poisson random variable based of i.i.d. observations $X_1, \ldots, X_n$. Is the estimator unbiased and consistent?

2. Caleb builds a particle detector and uses it to measure radiation from far stars. On any given day, the number of particles Y that hit the detector is conditionally distributed according to a Poisson distribution conditioned on parameter x. The parameter x is unknown and is modeled as the value of a random variable X, exponentially distributed with parameter μ as follows.

$$f_X(x) = \begin{cases} \mu e^{-\mu x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Then, the conditional PDF of the number of particles hitting the detector is,

$$p_{Y|X}(y \mid x) = \begin{cases} \frac{e^{-x}x^y}{y!} & \text{y = 0, 1, 2, ...} \\ 0 & \text{otherwise} \end{cases}$$

   (a) Find the MAP estimate of X from the observed particle count y.

   (b) Our goal is to find the conditional expectation estimator for X from the observed particle count y.

      i. Show that the posterior probability distribution for X given Y is of the form

$$f_{X|Y}(x \mid y) = \frac{\lambda^{y+1}}{y!} x^y e^{-\lambda x}, \quad x > 0$$

      and find the parameter λ. You may find the following equality useful (it is obviously true if the equation above describes a true PDF):

$$\int_0^\infty a^{y+1} x^y e^{-ax}\, dx = y! \quad \text{for any } a > 0$$

      ii. Find the conditional expectation estimate of X from the observed particle count y. *Hint:* you might want to express $x f_{X|Y}(x \mid y)$ in terms of $f_{X|Y}(x \mid y+1)$.

   (c) Compare the two estimators you constructed in part (a) and part (b).

3. Consider a Bernoulli process $X_1, X_2, X_3, \ldots$ with unknown probability of success q. 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 $\{t_1, t_2, t_3, \ldots\}$. In problem set 10, we solved the problem using Bayesian inference. Our focus here will be on classical estimation.

We assume that q is an unknown parameter in the interval $(0, 1]$. Denote the true parameter by q^*. Denote by $\widehat{Q}_k$ the maximum likelihood estimate (MLE) of q given k recordings, $T_1 = t_1, \ldots, T_k = t_k$.

(a) Compute $\widehat{Q}_k$. Is this different from the MAP estimate you found in problem set 10?

(b) Show that for all $\epsilon > 0$

$$\lim_{k\to\infty} \mathbf{P}\left(\left|\frac{1}{\widehat{Q}_k} - \frac{1}{q^*}\right| > \epsilon\right) = 0$$

(c) Assume $q^* \geq 0.5$. Give a lower bound on k such that

$$\mathbf{P}\left(\left|\frac{1}{\widehat{Q}_k} - \frac{1}{q^*}\right| \leq 0.1\right) \geq 0.95$$

4. A body at temperature θ radiates photons at a given wavelength. This problem will have you estimate θ, which is fixed but unknown. The PMF for the number of photons K in a given wavelength range and a fixed time interval of one second is given by,

$$p_K(k;\theta) = \frac{1}{Z(\theta)} e^{-\frac{k}{\theta}}, k = 0, 1, 2, \ldots$$

$Z(\theta)$ is a normalization factor for the probability distribution (the physicists call it the partition function). You are given the task of determining the temperature of the body to two significant digits by photon counting in non-overlapping time intervals of duration one second. The photon emissions in non-overlapping time intervals are statistically independent from each other.

(a) Determine the normalization factor $Z(\theta)$.

(b) Compute the expected value of the photon number measured in any 1 second time interval, $\mu_K = \mathbf{E}_\theta[K]$, and its variance, $\text{var}_\theta(K) = \sigma_K^2$.

(c) You count the number k_i of photons detected in n non-overlapping 1 second time intervals. Find the maximum likelihood estimator, $\hat{\Theta}_n$, for temperature Θ. Note, it might be useful to introduce the average photon number $s_n = \frac{1}{n}\sum_{i=1}^{n} k_i$. In order to keep the analysis simple we assume that the body is hot, i.e. $\theta \gg 1$.
You may use the approximation: $\frac{1}{e^{\frac{1}{\theta}} - 1} \approx \theta$ for $\theta \gg 1$.

In the following questions we wish to estimate the mean of the photon count in a one second time interval using the estimator $\hat{K}$, which is given by,

$$\hat{K} = \frac{1}{n}\sum_{i=1}^{n} K_i.$$

(d) Find the number of samples n for which the noise to signal ratio for $\hat{K}$, (i.e.,$\frac{\sigma_{\hat{K}}}{\mu_{\hat{K}}}$), is 0.01.

(e) Find the 95% confidence interval for the mean photon count estimate for the situation in part (d). (You may use the central limit theorem.)

5. The RandomView window factory produces window panes. After manufacturing, 1000 panes were loaded onto a truck. The weight W_i of the i-th pane (in pounds) on the truck is modeled as a random variable, with the assumption that the W_i's are independent and identically distributed.

(a) Assume that the measured weight of the load on the truck was 2340 pounds, and that $\text{var}(W_i) \leq 4$. Find an approximate 95 percent confidence interval for $\mu = \mathbf{E}[W_i]$, using the Central Limit Theorem.

(b) Now assume instead that the random variables W_i are i.i.d., with an exponential distribution with parameter $\theta > 0$, i.e., a distribution with PDF

$$f_W(w; \theta) = \theta e^{-\theta w}.$$

What is the maximum likelihood estimate of θ, given that the truckload has weight 2340 pounds?

6. Given the five data pairs (x_i, y_i) in the table below,

x	0.8	2.5	5	7.3	9.1
y	-2.3	20.9	103.5	215.8	334

we want to construct a model relating x and y. We consider a linear model

$$Y_i = \theta_0 + \theta_1 x_i + W_i, \qquad i = 1, \cdots, 5,$$

and a quadratic model

$$Y_i = \beta_0 + \beta_1 x_i^2 + V_i, \qquad i = 1, \cdots, 5.$$

where W_i and V_i represent additive noise terms, modeled by independent normal random variables with mean zero and variance σ_1^2 and σ_2^2, respectively.

(a) Find the ML estimates of the linear model parameters.

(b) Find the ML estimates of the quadratic model parameters.

Note: You may use the regression formulas and the connection with ML described in pages 478-479 of the text. However, the regression material is outside the scope of the final.

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

Problem Set 11 Solutions

1. Check book solutions on Stellar.

2. (a) To find the MAP estimate, we need to find the value x that maximizes the conditional density $f_{X|Y}(x \mid y)$ by taking its derivative and setting it to 0.

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{p_{Y|X}(y \mid x) \cdot f_X(x)}{p_Y(y)} \\ &= \frac{e^{-x}x^y}{y!} \cdot \mu e^{-\mu x} \cdot \frac{1}{p_Y(y)} \\ &= \frac{\mu}{y! p_Y(y)} \cdot e^{-(\mu+1)x} x^y \end{aligned}$$

$$\begin{aligned} \frac{d}{dx} f_{X|Y}(x \mid y) &= \frac{d}{dx}\left(\frac{\mu}{y! p_Y(y)} \cdot e^{-(\mu+1)x} x^y\right) \\ &= \frac{\mu}{y! p_Y(y)} x^{y-1}\, e^{-(\mu+1)x}\, (y - x(\mu+1)) \end{aligned}$$

Since the only factor that depends on x which can take on the value 0 is $(y - x(\mu+1))$, the maximum is achieved at

$$\hat{x}_{\text{MAP}}(y) = \frac{y}{1+\mu}$$

It is easy to check that this value is indeed maximum (the first derivative changes from positive to negative at this value).

(b) i. To show the given identity, we need to use Bayes' rule. We first compute the denominator, $p_Y(y)$

$$\begin{aligned} p_Y(y) &= \int_0^\infty \frac{e^{-x}x^y}{y!} \mu e^{-\mu x}\, dx \\ &= \frac{\mu}{y!\,(1+\mu)^{y+1}} \int_0^\infty (1+\mu)^{y+1} x^y e^{-(1+\mu)x}\, dx \\ &= \frac{\mu}{(1+\mu)^{y+1}} \end{aligned}$$

Then, we can substitute into the equation we had derived in part (a)

$$\begin{aligned} f_{X|Y}(x \mid y) &= \frac{\mu}{y! p_Y(y)} x^y e^{-(\mu+1)x} \\ &= \frac{\mu}{y!} \frac{(1+\mu)^{y+1}}{\mu} x^y e^{-(\mu+1)x} \\ &= \frac{(1+\mu)^{y+1}}{y!} x^y e^{-(\mu+1)x} \end{aligned}$$

Thus, $\lambda = 1 + \mu$.

ii. We first manipulate $xf_{X|Y}(x \mid y)$:

$$\begin{aligned} xf_{X|Y}(x \mid y) &= \frac{(1+\mu)^{y+1}}{y!} x^{y+1} e^{-(\mu+1)x} \\ &= \frac{y+1}{1+\mu} \frac{(1+\mu)^{y+2}}{(y+1)!} x^{y+1} e^{-(\mu+1)x} \\ &= \frac{y+1}{1+\mu} f_{X|Y}(x \mid y+1) \end{aligned}$$

Now we can find the conditional expectation estimator:

$$\begin{aligned} \hat{x}_{\mathrm{CE}}(y) &= \mathbf{E}[X|Y=y] = \int_0^\infty x f_{X|Y}(x \mid y)\, dx \\ &= \int_0^\infty \frac{y+1}{1+\mu} f_{X|Y}(x \mid y+1)\, dx = \frac{y+1}{1+\mu} \end{aligned}$$

(c) The conditional expectation estimator is always higher than the MAP estimator by $\frac{1}{1+\mu}$.

3. (a) The likelihood function is

$$\prod_{i=1}^{k} P_{T_i}(T_i = t_i \mid Q = q) = q^k (1-q)^{\sum_i^k t_i - k}.$$

To maximize the above probability we set its derivative with respect to q to zero

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

or equivalently

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

which yields $\widehat{Q}_k = \frac{k}{\sum_{i=1}^k t_i}$. This is not different from the MAP estimate found before. Since the MAP estimate is calculated using a uniform prior, the likelihood function is a 'scaled' version of posterior probability and they can be maximized at the same value of q.

(b) Since $\dfrac{1}{\widehat{Q}_k} = \dfrac{\sum_{i=1}^k T_i}{k}$, and that each T_i is independent identically distributed, it follows that $\dfrac{1}{\widehat{Q}_k}$ is actually a sample mean estimator. The weak law of large numbers says that, when the number of samples increases to infinity, the sample mean estimator converges to the actual mean, which is $\frac{1}{q^*}$ in this case. So we can write the limit of probability as

$$\lim_{k\to\infty} \mathbf{P}\left(\left| \frac{1}{\widehat{Q}_k} - \frac{1}{q^*} \right| > \epsilon \right) = \lim_{k\to\infty} \mathbf{P}\left(\left| \frac{\sum_{i=1}^k T_i}{k} - \mathbf{E}[T_1] \right| > \epsilon \right) = 0.$$

(c) Chebyshev inequality states that

$$\mathbf{P}\left(\left|\frac{\sum_{i=1}^{k} T_i}{k} - \mathbf{E}[T_1]\right| \geq \epsilon\right) \leq \frac{\text{var}(T_1)}{k\epsilon^2}.$$

So we have

$$\mathbf{P}\left(\left|\frac{1}{\widehat{Q}_k} - \frac{1}{q^*}\right| \leq 0.1\right) = \mathbf{P}\left(\left|\frac{\sum_{i=1}^{k} T_i}{k} - \frac{1}{q^*}\right| \leq 0.1\right)$$
$$= 1 - \mathbf{P}\left(\left|\frac{\sum_{i=1}^{k} T_i}{k} - \mathbf{E}[T_1]\right| \geq 0.1\right) \geq 1 - \frac{\text{var}(T_1)}{k * 0.1^2}$$

To ensure the above probability to be greater than 0.95, we need that

$$1 - \frac{\text{var}(T_1)}{k * 0.1^2} = 1 - \frac{\frac{1-q}{q^2}}{k * 0.1^2} \geq 0.95,$$

or

$$k \geq 2000\text{var}(T_1) = 2000\frac{1-q}{q^2}$$

The number of observations k needed depends on the variance of T_1. For q close to 1, the variance is close to 0, and the required number of observations is very small (close to 0). For $q = 1/2$, the variance is maximum ($\text{var}(T_1) = 2$), and we require $k = 4000$. Thus, to guarantee the required accuracy and confidence for all q, we need that,

$$k \geq 4000.$$

4. (a) Normalization of the distribution requires:

$$1 = \sum_{k=0}^{\infty} p_K(k;\theta) = \sum_{k=0}^{\infty} \frac{e^{-\frac{k}{\theta}}}{Z(\theta)} = \frac{1}{Z(\theta)} \sum_{k=0}^{\infty} e^{-\frac{k}{\theta}} = \frac{1}{Z(\theta) \cdot (1 - e^{-\frac{1}{\theta}})},$$

so $Z(\theta) = \frac{1}{1-e^{-\frac{1}{\theta}}}$.

(b) Rewriting $p_K(k;\theta)$ as:

$$p_K(k;\theta) = \left(e^{-\frac{1}{\theta}}\right)^k \left(1 - e^{-\frac{1}{\theta}}\right), \qquad k = 0, 1, \ldots$$

the probability distribution for the photon number is a geometric probability distribution with probability of success $p = 1 - e^{-\frac{1}{\theta}}$, and it is shifted with 1 to the left since it starts with $k = 0$. Therefore the photon number expectation value is

$$\mu_K = \frac{1}{p} - 1 = \frac{1}{1 - e^{-\frac{1}{\theta}}} - 1 = \frac{1}{e^{\frac{1}{\theta}} - 1}$$

and its variance is

$$\sigma_K^2 = \frac{1-p}{p^2} = \frac{e^{-\frac{1}{\theta}}}{(1 - e^{-\frac{1}{\theta}})^2} = \mu_K^2 + \mu_K.$$

(c) The joint probability distribution for the k_i is

$$p_K(k_1, ..., k_n; \theta) = \frac{1}{Z(\theta)^n} \Pi_{i=1}^n e^{-k_i/\theta} = \frac{1}{Z(\theta)^n} e^{-\frac{1}{\theta}\sum_{i=1}^n k_i}.$$

The log likelihood is $-n \cdot \log Z(\theta) - 1/\theta \sum_{i=1}^n k_i$.
We find the maxima of the log likelihood by setting the derivative with respect to the parameter θ to zero:

$$\frac{d}{d\theta} \log p_K(k_1, ..., k_n; \theta) = -n \cdot \frac{e^{-\frac{1}{\theta}}}{\theta^2(1 - e^{-\frac{1}{\theta}})} + \frac{1}{\theta^2} \sum_{i=1}^n k_i = 0$$

or

$$\frac{1}{e^{\frac{1}{\theta}} - 1} = \frac{1}{n} \sum_{i=1}^n k_i = s_n.$$

For a hot body, $\theta \gg 1$ and $\frac{1}{e^{\frac{1}{\theta}} - 1} \approx \theta$, we obtain

$$\theta \approx \frac{1}{n} \sum_{i=1}^n k_i = s_n.$$

Thus the maximum likelihood estimator $\hat{\Theta}_n$ for the temperature is given in this limit by the sample mean of the photon number

$$\hat{\Theta}_n = \frac{1}{n} \sum_{i=1}^n K_i.$$

(d) According to the central limit theorem, the sample mean approaches for large n a Gaussian distribution with standard deviation our root mean square error

$$\sigma_{\hat{\Theta}_n} = \frac{\sigma_K}{\sqrt{n}}.$$

To allow only for 1% relative root mean square error in the temperature, we need $\frac{\sigma_K}{\sqrt{n}} < 0.01\mu_K$. With $\sigma_K^2 = \mu_K^2 + \mu_K$ it follows that

$$\sqrt{n} > \frac{\sigma_K}{0.01\mu_K} = 100 \frac{\sqrt{\mu_K^2 + \mu_K}}{\mu_K} = 100\sqrt{1 + \frac{1}{\mu_K}}.$$

In general, for large temperatures, i.e. large mean photon numbers $\mu_K \gg 1$, we need about 10,000 samples.

(e) The 95% confidence interval for the temperature estimate for the situation in part (d), i.e.

$$\sigma_{\hat{\Theta}_n} = \frac{\sigma_K}{\sqrt{n}} = 0.01\mu_K,$$

is

$$[\hat{K} - 1.96\sigma_{\hat{K}}, \hat{K} + 1.96\sigma_{\hat{K}}] = [\hat{K} - 0.0196\mu_K, \hat{K} + 0.0196\mu_K].$$

5. (a) The sample mean estimator $\hat{\Theta}_n = \frac{W_1+\cdots+W_n}{n}$ in this case is

$$\hat{\Theta}_{1000} = \frac{2340}{1000} = 2.34.$$

From the standard normal table, we have $\Phi(1.96) = 0.975$, so we obtain

$$\mathbf{P}\left(\frac{|\hat{\Theta}_{1000} - \mu|}{\sqrt{\text{var}(W_i)/1000}} \leq 1.96\right) \approx 0.95.$$

Because the variance is less that 4, we have

$$\mathbf{P}\left(\hat{\Theta}_{1000} - \mu \leq 1.96\sqrt{\text{var}(W_i)/1000}\right) \leq \mathbf{P}\left(\hat{\Theta}_{1000} - \mu \leq 1.96\sqrt{4/1000}\right),$$

and letting the right-hand side of the above equation ≈ 0.95 gives a 95% confidence, i.e.,

$$\left[\hat{\Theta}_{1000} - 1.96\sqrt{4/1000}, \hat{\Theta}_{1000} + 1.96\sqrt{4/1000}\right] = \left[\hat{\Theta}_{1000} - 0.124, \hat{\Theta}_{1000} + 0.124\right] = [2.216, 2.464]$$

(b) The likelihood function is

$$f_W(w;\theta) = \prod_{i=1}^{n} f_{W_i}(w_i;\theta) = \prod_{i=1}^{n} \theta e^{-\theta w_i},$$

And the log-likelihood function is

$$\log f_W(w;\theta) = n\log\theta - \theta\sum_{i=1}^{n} w_i,$$

The derivative with respect to θ is $\frac{n}{\theta} - \sum_{i=1}^{n} w_i$,and by setting it to zero, we see that the maximum of $\log f_W(w;\theta)$ over $\theta \geq 0$ is attained at $\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} w_i}$. The resulting estimator is

$$\hat{\Theta}_n^{mle} = \frac{n}{\sum_{i=1}^{n} W_i}.$$

In this case,

$$\hat{\Theta}_n^{mle} = \frac{1000}{2340} = 0.4274.$$

6. (a) Using the regression formulas of Section 9.2, we have

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i - \bar{x})^2}, \qquad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x},$$

where

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = 4.94, \qquad \bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = 134.38.$$

The resulting ML estimates are

$$\hat{\theta}_1 = 40.53, \qquad \hat{\theta}_0 = -65.86.$$

(b) Using the same procedure as in part (a), we obtain

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{5}(x_i^2 - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{5}(x_i^2 - \bar{x})^2}, \qquad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1\bar{x},$$

where

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i^2 = 33.60, \qquad \bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = 134.38.$$

which for the given data yields

$$\hat{\theta}_1 = 4.09, \qquad \hat{\theta}_0 = -3.07.$$

Figure 1 shows the data points (x_i, y_i), $i = 1, \ldots, 5$, the estimated linear model

$$y = 40.53x - 65.86,$$

and the estimated quadratic model

$$y = 4.09x^2 - 3.07.$$

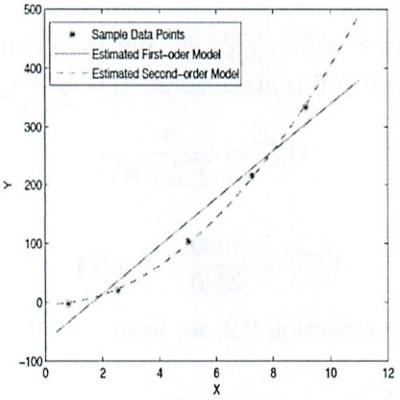


Figure 1: Regression Plot

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

Recitation 25
December 9, 2010
Based on Spring 10 Final exam

Question 2 An atom of the radioactive element Vestium decays to an atom of Hockfieldium after a time that is an exponential random variable with parameter λ. Hockfieldium is a stable element; i.e., it is not radioactive. Each radioactive atom decays independently of any other atoms.

(a) Suppose a box has n atoms of Vestium at time 0, where n is a positive integer. Let V be the remaining atoms of Vestium in the box at time t, where t is a positive real number. Find the PMF of V.

(b) An atom of Vestium can itself be the product of the radioactive decay of an atom of Grayon. The decay of any one atom of Grayon to an atom of Vestium occurs after a time that is an exponential random variable with parameter μ.

Suppose a box initially contains two atoms of Grayon and nothing else. Find the expected time until the box is no longer radioactive, i.e., it contains neither Grayon nor Vestium—only Hockfieldium.

Question 4 **Breaking a stick more than twice.** We start with a stick of length ℓ. We break at a point which is chosen randomly and uniformly over its length, and keep the piece that contains the left end of the stick. We then repeat the same process several times on the piece that we were left with. Denote by X_n the length of the piece we are left with after breaking n times.

(a) Find $\mathbf{E}[X_n]$.

(b) After breaking the stick n times, we randomly pick one of the $n+1$ pieces, each of the pieces being equally likely to be picked. Calculate the expected length of the chosen piece.

(c) Does the sequence $X_1, X_2, \ldots$ converge in probability to a number? If so, to what value? Prove.

Question 5 Let W_1, W_2, and W_3 be independent, continuous random variables each uniformly distributed over $[0, 1]$. Let $X = W_1 + W_2$ and $Y = X + W_3$.

(a) Find the linear least mean squares (LLMS) estimator of X from Y.

(b) Find the maximum a posteriori probability (MAP) estimator of X from Y.

Recetation 25

12/9

-Focus on 2nd half of course

-this recitation samples everything except Markov

2. Graysom → Vestium → Hoch Fiedium

G atom → V atom after decay time $\sim \exp(\mu)$

$V \rightarrow H \sim \exp(\lambda)$

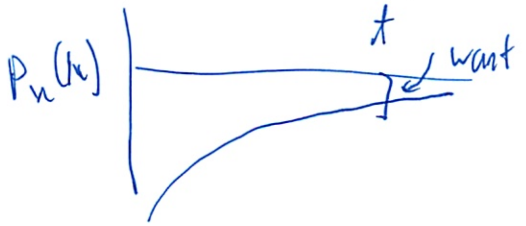
a) $\boxed{n \text{ Vatoms}} \xrightarrow[\text{time } t]{\text{After fixed}} \boxed{k \text{ V atoms}}$

Want PDF of k

- each one decays ind.
- if it does <u>not</u> decay = sucess

$p_K(k) \sim$ Binomial(p)

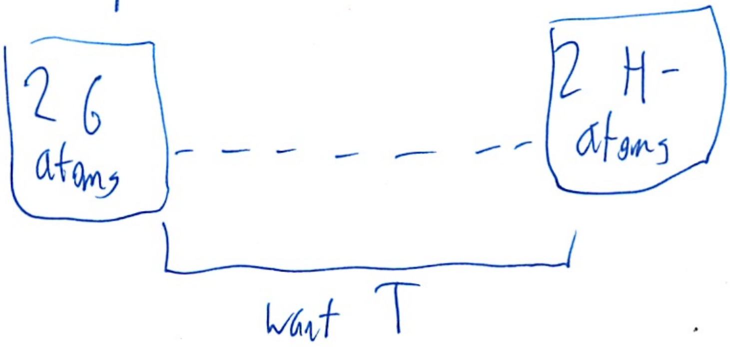
p = P(No decay by time t) $= 1 - e^{-\lambda t}$



$$= \binom{n}{k}\left(e^{-\lambda t}\right)^k\left(1 - e^{-\lambda t}\right)^{n-k} \quad k = 0, 1, \ldots, n$$

②

b) More poisson like

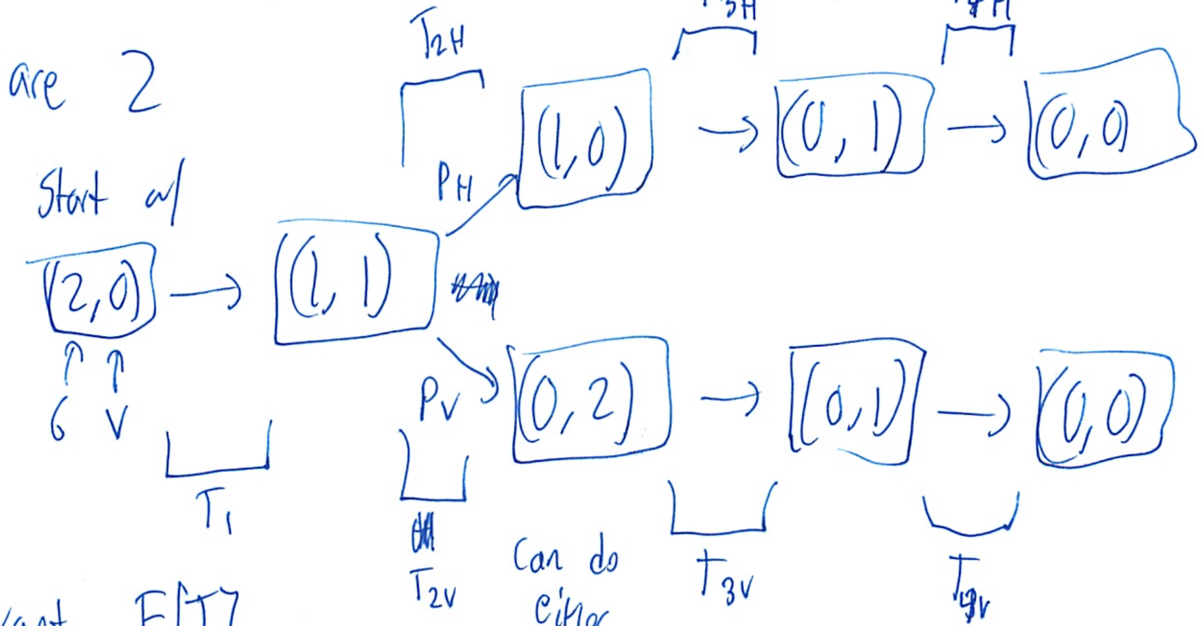


If only 1

$$\frac{1}{\mu} + \frac{1}{\lambda}$$

G→V V→H

But are 2



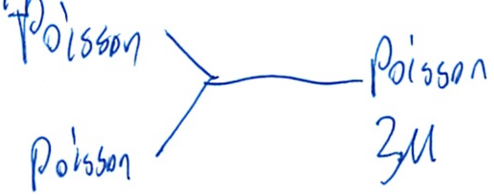
Want $E[T]$

- calculate T upper
  lower
- then expected value theorm

$$E[T] = P_H \cdot E[\text{time of upper bound}] + P_V \cdot E[\text{time of lower bound}]$$

③

$E[T_1] =$



What is time of 1st arrival in 2μ Poisson merged process?

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

$$E[T_{2H}] = \frac{1}{\lambda}$$

$E[T_{3H}] = \frac{1}{\mu}$ time for G→V

$E[T_{4H}] = \frac{1}{\lambda}$ time V→H

$$E[T_{2V}] = \frac{1}{\mu}$$

$E[T_{3V}] = \frac{1}{2\lambda}$ ← from merged 2λ process – either one can leave

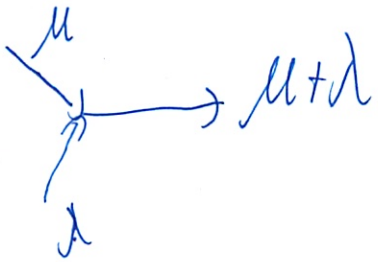
$$E[T_{4V}] = \frac{1}{\lambda}$$

$$E[T] = P_H \cdot \left(\frac{1}{2\mu} + \frac{1}{\lambda} + \frac{1}{\mu} + \frac{1}{\lambda}\right) + P_V \cdot \left(\frac{1}{2\mu} + \frac{1}{\mu} + \frac{1}{2\lambda} + \frac{1}{\lambda}\right)$$

Note the two branches are different bottom branch is faster

(this all seems so easy – can I think of it?)

④



$$P_{H} = \frac{\lambda}{\mu+\lambda}$$

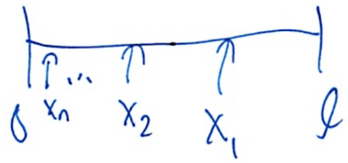
$$P_{V} = \frac{\mu}{\mu+\lambda}$$

$$E[T] = \frac{\lambda}{\mu+\lambda} \cdot \left(\frac{1}{2\mu} + \frac{1}{\lambda} + \frac{1}{\lambda} + \frac{1}{\lambda}\right) + \frac{\mu}{\mu+\lambda}\left(\frac{1}{2\mu} + \frac{1}{\mu} + \frac{1}{2\lambda} + \frac{1}{\lambda}\right)$$

---

4. Stick breaking problem

(try to get intuition to do it - don't just memorize example)



X_n = lenght of left most piece after n breaks

$E[X_n] = ?$

$E[X_1] = \frac{\ell}{2}$ ← since break according to uniform dist

$E[X_2]$ = conditioning + law of iterated expectation

$= E[E[X_n \mid X_{n-1}]]$

$= E[X_{n-1}/2]$

⑤

$$= \frac{E[X_{n-1}]}{2}$$

$E[X_n] =$ generalize

$$= \frac{\ell}{2^n}$$

b) After n breaks, of lenghts $L_0, L_1, \ldots, L_n$
have n+1 pieces

M: Lenght of randomly chosen piece

↳ prob $\frac{1}{n+1}$

$$E[M] \overset{?}{=} \sum_{i=0}^{n} \underbrace{P(\text{ith piece chosen})}_{\uparrow \frac{1}{n+1}} \underbrace{E[M \mid \text{ith piece chosen}]}_{\uparrow E[L_i]}$$

$$= \frac{1}{n+1} \sum_{i=1}^{n} E[L_i]$$

$$= \frac{1}{n+1} E\left[\sum_{i=0}^{n} L_i\right]$$

$= \frac{\ell}{n+1}$ ← all of the pieces sum up to ℓ

⑥

c) As n get larger, L_n gets smaller + smaller + converges to 0

Dees $\{X_n\}$ converge in prob and to what limit?

- to 0

Does it converge?

- fancy math
- but straightforward - only 1 path for proof

For any ϵ, find $P(|X_n - 0| \geq \epsilon)$

>0

Show converges $\to 0$ as $n \to \infty$

Markov inequality

$$P(X_n \geq \epsilon) \leq \frac{E[X_n]}{\epsilon}$$

$$= \frac{1}{\epsilon \cdot 2^n}$$

$\to 0$ as $n \to \infty$

Q.E.D

Proved $\to 0$ as $n \to \infty$ for any $\epsilon > 0$

5. Inference problem

Baysian - MAP, LMS

Classical - ML

Have 3 RV $W_1, W_2, W_3 \rightarrow$ iid

Uniformly distributed

$X = W_1 + W_2$

$Y = X + W_3$ ⤶ X, Y correlated

If know Y, can make an infrence about X

$$\hat{X}_{LMS} = E[X] + \frac{Cov(X,Y)}{var(Y)}(Y - E[Y])$$

Find these values + plug in

$$E[X] = E[W_1] + E[W_2]$$
$$= \frac{1}{2} + \frac{1}{2}$$
$$= 1$$

$$E[Y] = E[X] + E[W_3]$$
$$= 1 + \frac{1}{2}$$

ind. does not matter

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

⑧

$Var(x) =$ sum of two ind RV = sum of var

$= Var(W_1) + var(w_2)$

$= \frac{1}{12} + \frac{1}{12}$

$= \frac{1}{6}$

$Var(y) = var(X) + var(w_3)$

$= \frac{1}{6} + \frac{1}{12}$

$= \frac{1}{4}$

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

need cross
2nd moment

$E[XY] = E[X(X + W_3)]$

= both ind so $E[XY] = E[X]\,E[Y]$

$= E[X^2] + E[X]\,E[W_3]$

$= Var(x) + (E[x])^2 + E[X]\,E[W_3]$

$= \frac{1}{6} + 1^2 + 1 \cdot \frac{1}{2}$

$= \frac{5}{3}$

⑨

$$\text{Cov}(X,Y) = \frac{5}{3} - 1 \cdot \frac{2}{3}$$

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

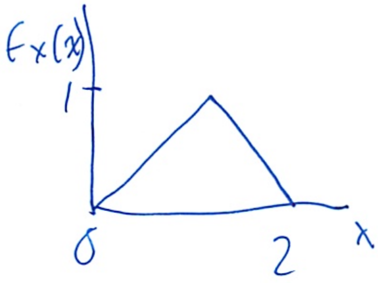
$$\hat{X}_{LLMS} = 1 + \frac{1/6}{1/4}\left(Y - \frac{3}{2}\right)$$

$$= 1 + \frac{2}{3}\left(Y - \frac{3}{2}\right)$$

$$= \frac{2}{3} Y$$

b) MAP estimator of X based on Y

$f_X(x)$ = triangular density though lenghty convoltion process



~~MAP is max posterior here = $\hat{X}_{MAP}$~~

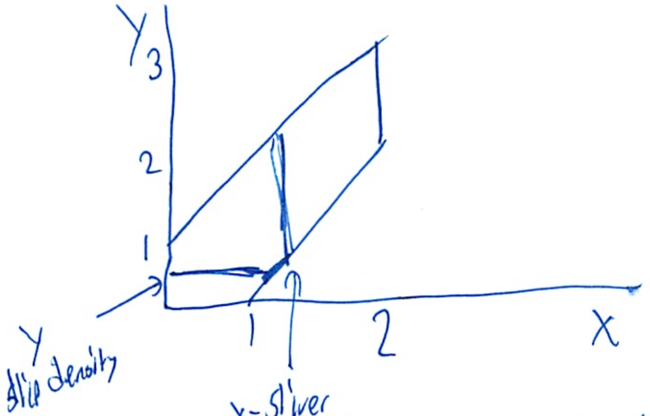
MAP → $\max_x f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$ ← can calcdate easily

denominator does not matter

$$= \max_x f_{X,Y}(x,y)$$

⑩

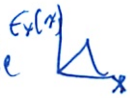
$Y = X + \underbrace{[0,1]}_{\text{Uniform}}$



joint PDF

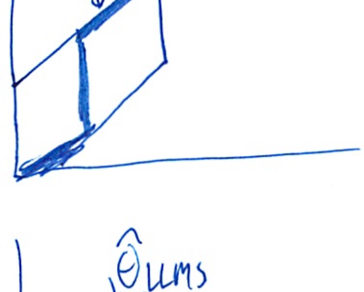
x= sliver
more dense
here, in middle

— see the triangle

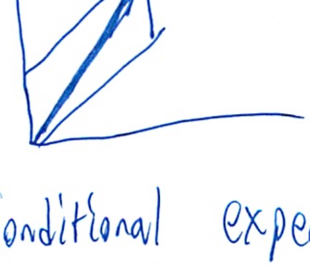


Want density along this slip.

$$\hat{X}_{MAP} = \begin{cases} 1 & \text{if } 1 \le y \le 2 \\ Y & \text{if } 0 \le y \le 1 \\ Y-1 & \text{if } 2 \le y \le 3 \end{cases}$$



Where is joint maximized?



Conditional expectation estimator

— integrate along slice —

⑪

Must do calculation

In this case same as LLMS

^not true in general

9.5 Summary

Have not discussed

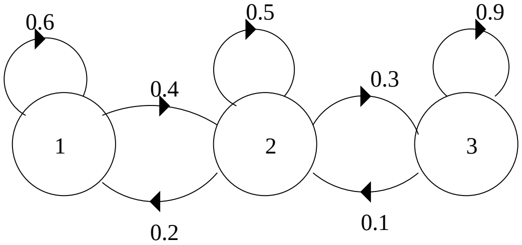
- ~~An~~ estimation in time-varying env
- non parametric estimation
- further developments in linear + non-linear regression
- methods for designing statistical experiments
- methods for validating conclusions of stat. study
- Computational methods
- etc

end of book!

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

6.041/6.431 Fall 2010 Final Exam Solutions
Wednesday, December 15, 9:00AM - 12:00noon.

Problem 1. (32 points) Consider a Markov chain $\{X_n; n = 0, 1, \ldots\}$, specified by the following transition diagram.



1. **(4 points)** Given that the chain starts with $X_0 = 1$, find the probability that $X_2 = 2$.

   Solution: The two-step transition probability is:

$$\begin{aligned} r_{12}(2) &= p_{11} \cdot p_{12} + p_{12} \cdot p_{22} \\ &= 0.6 \cdot 0.4 + 0.4 \cdot 0.5 \\ &= 0.44. \end{aligned}$$

2. **(4 points)** Find the steady-state probabilities π_1, π_2, π_3 of the different states.

   Solution: We set up the balance equations of a birth-death process and the normalization equation as such:

$$\begin{aligned} \pi_1 p_{12} &= \pi_2 p_{21} \\ \pi_2 p_{23} &= \pi_3 p_{32} \\ \pi_1 + \pi_2 + \pi_3 &= 1. \end{aligned}$$

   Solving the system of equations yields the following steady state probabilities:

$$\begin{aligned} \pi_1 &= 1/9 \\ \pi_2 &= 2/9 \\ \pi_3 &= 6/9. \end{aligned}$$

   In case you did not do part 2 correctly, in **all** *subsequent parts of this problem you can just use the symbols π_i: you do not need to plug in actual numbers.*

3. **(4 points)** Let $Y_n = X_n - X_{n-1}$. Thus, $Y_n = 1$ indicates that the nth transition was to the right, $Y_n = 0$ indicates it was a self-transition, and $Y_n = -1$ indicates it was a transition to the left. Find $\lim\limits_{n\to\infty} \mathbf{P}(Y_n = 1)$.

   Solution: Using the total probability theorem and steady state probabilities,

$$\begin{aligned} \lim_{n\to\infty} \mathbf{P}(Y_n = 1) &= \sum_{i=1}^{3} \pi_i \cdot \mathbf{P}(Y_n = 1 \mid X_{n-1} = i) \\ &= \pi_1 p_{12} + \pi_2 p_{23} \\ &= 1/9. \end{aligned}$$

4. **(4 points)** Is the sequence Y_n a Markov chain? Justify your answer.

   Solution: No. Assume the Markov process is in steady state. To satisfy the Markov property,

$$\mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1, Y_{n-2} = 1) = \mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1).$$

   For large n,

$$\mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1, Y_{n-2} = 1) = 0,$$

   since it is not possible to move upwards 3 times in a row. However in steady state,

$$\begin{aligned}\mathbf{P}(Y_n = 1 \mid Y_{n-1} = 1) &= \frac{\mathbf{P}(\{Y_n = 1\} \cap \{Y_{n-1} = 1\})}{\mathbf{P}(Y_{n-1} = 1)} \\ &= \frac{\pi_1 p_{12} p_{23}}{\pi_1 p_{12} + \pi_2 p_{23}} \\ &\neq 0.\end{aligned}$$

   Therefore, the sequence Y_n is not a Markov chain.

5. **(4 points)** Given that the nth transition was a transition to the right ($Y_n = 1$), find the probability that the previous state was state 1. (You can assume that n is large.)

   Solution: Using Bayes' Rule,

$$\begin{aligned}\mathbf{P}(X_{n-1} = 1 \mid Y_n = 1) &= \frac{\mathbf{P}(X_{n-1} = 1)\mathbf{P}(Y_n = 1 \mid X_{n-1} = 1)}{\sum_{i=1}^{3} \mathbf{P}(X_{n-1} = i)\mathbf{P}(Y_n = 1 \mid X_{n-1} = i)} \\ &= \frac{\pi_1 p_{12}}{\pi_1 p_{12} + \pi_2 p_{23}} \\ &= 2/5.\end{aligned}$$

6. **(4 points)** Suppose that $X_0 = 1$. Let T be defined as the first *positive time* at which the state is again equal to 1. Show how to find $\mathbf{E}[T]$. (It is enough to write down whatever equation(s) needs to be solved; you do not have to actually solve it/them or to produce a numerical answer.)

   Solution: In order to find the the mean recurrence time of state 1, the mean first passage times to state 1 are first calculated by solving the following system of equations:

$$\begin{aligned}t_2 &= 1 + p_{22}t_2 + p_{23}t_3 \\ t_3 &= 1 + p_{32}t_2 + p_{33}t_3.\end{aligned}$$

   The mean recurrence time of state 1 is then $t_1^* = 1 + p_{12}t_2$.

   Solving the system of equations yields $t_2 = 20$ and $t_3 = 30$ and $t_1^* = 9$.

7. **(4 points)** Does the sequence $X_1, X_2, X_3, \ldots$ converge in probability? If yes, to what? If not, just say "no" without explanation.

   Solution: No.

8. **(4 points)** Let $Z_n = \max\{X_1, \ldots, X_n\}$. Does the sequence $Z_1, Z_2, Z_3, \ldots$ converge in probability? If yes, to what? If not, just say "no" without explanation.

   Solution: Yes. The sequence converges to 3 in probability.

   For the original markov chain, states $\{1, 2, 3\}$ form one single recurrent class. Therefore, the Markov process will eventually visit each state with probability 1. In this case, the sequence Z_n will, with probability 1, converge to 3 once X_n visits 3 for the first time.

Problem 2. (68 points) Alice shows up at an Athena cluster at time zero and spends her time exclusively in typing emails. The times that her emails are sent are a Poisson process with rate λ_A per hour.

1. **(3 points)** What is the probability that Alice sent exactly three emails during the time interval $[1, 2]$?

   Solution: The number of emails Alice sends in the interval $[1, 2]$ is a Poisson random variable with parameter λ_A. So we have:

$$\mathbf{P}(3, 1) = \frac{{\lambda_A}^3 e^{-\lambda_A}}{3!}.$$

2. Let Y_1 and Y_2 be the times at which Alice's first and second emails were sent.

   (a) **(3 points)** Find $\mathbf{E}[Y_2 \mid Y_1]$.
   Solution: Define T_2 as the second inter-arrival time in Alice's Poisson process. Then:

$$Y_2 = Y_1 + T_2$$

$$\mathbf{E}[Y_2 \mid Y_1] = \mathbf{E}[Y_1 + T_2 \mid Y_1] = Y_1 + \mathbf{E}[T_2] = Y_1 + 1/\lambda_A.$$

   (b) **(3 points)** Find the PDF of Y_1^2.
   Solution: Let $Z = Y_1^2$. Then we first find the CDF of Z and differentiate to find the PDF of Z:

$$F_Z(z) = \mathbf{P}(Y_1^2 \le z) = \mathbf{P}(-\sqrt{z} \le Y_1 \le \sqrt{z}) = \begin{cases} 1 - e^{-\lambda_A \sqrt{z}} & z \ge 0 \\ 0 & z < 0. \end{cases}$$

$$\begin{aligned} f_Z(z) = \frac{dF_Z(z)}{dz} &= \lambda_A e^{-\lambda_A \sqrt{z}} \left(\frac{1}{2} z^{-1/2} \right) \qquad (z \ge 0) \\ f_Z(z) &= \begin{cases} \frac{\lambda_A}{2\sqrt{z}} e^{-\lambda_A \sqrt{z}} & z \ge 0 \\ 0 & z < 0. \end{cases} \end{aligned}$$

   (c) **(3 points)** Find the joint PDF of Y_1 and Y_2.
   Solution:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= f_{Y_1}(y_1) f_{Y_2|Y_1}(y_2|y_1) \\ &= f_{Y_1}(y_1) f_{T_2}(y_2 - y_1) \\ &= \lambda_A e^{-\lambda_A y_1} \lambda_A e^{-\lambda_A (y_2 - y_1)} \qquad y_2 \ge y_1 \ge 0 \\ &= \begin{cases} \lambda_A^2 e^{-\lambda_A y_2} & y_2 \ge y_1 \ge 0 \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

3. You show up at time 1 and you are told that Alice has sent exactly one email so far. (Only give answers here, no need to justify them.)

   (a) **(3 points)** What is the conditional expectation of Y_2 given this information?
   Solution: Let A be the event {exactly one arrival in the interval [0,1]}. Looking forward from time $t = 1$, the time until the next arrival is simply an exponential random variable (T). So,

$$\mathbf{E}[Y_2 \mid A] = 1 + \mathbf{E}[T] = 1 + 1/\lambda_A.$$

   (b) **(3 points)** What is the conditional expectation of Y_1 given this information?
   Solution: Given A, the times in this interval are equally likely for the arrival Y_1. Thus,

$$\mathbf{E}[Y_1 \mid A] = 1/2.$$

4. Bob just finished exercising (without email access) and sits next to Alice at time 1. He starts typing emails at time 1, and fires them according to an independent Poisson process with rate λ_B.

   (a) **(5 points)** What is the PMF of the total number of emails sent by the two of them together during the interval $[0, 2]$?
   Solution: Let K be the total number of emails sent in $[0, 2]$. Let K_1 be the total number of emails sent in $[0, 1)$, and let K_2 be the total number of emails sent in $[1, 2]$. Then $K = K_1 + K_2$ where K_1 is a Poisson random variable with parameter λ_A and K_2 is a Poisson random variable with parameter $\lambda_A + \lambda_B$ (since the emails sent by both Alice and Bob after time $t = 1$ arrive according to the merged Poisson process of Alice's emails and Bob's emails). Since K is the sum of independent Poisson random variables, K is a Poisson random variable with parameter $2\lambda_A + \lambda_B$. So K has the distribution:

$$p_K(k) = \frac{(2\lambda_A + \lambda_B)^k e^{-(2\lambda_A+\lambda_B)}}{k!} \qquad k = 0, 1, \ldots.$$

   (b) **(5 points)** What is the expected value of the total typing time associated with the email that Alice is typing at the time that Bob shows up? (Here, "total typing time" includes the time that Alice spent on that email both before and after Bob's arrival.)
   Solution: The total typing time Q associated with the email that Alice is typing at the time Bob shows up is the sum of S_0, the length of time between Alice's last email or time 0 (whichever is later) and time 1, and T_1, the length of time from 1 to the time at which Alice sends her current email. T_1 is exponential with parameter λ_A. and $S_0 = \min\{T_0, 1\}$, where T_0 is exponential with parameter λ_A.
   Then,

$$Q = S_0 + T_1 = \min\{T_0, 1\} + T_1$$

   and

$$\mathbf{E}[Q] = \mathbf{E}[S_0] + \mathbf{E}[T_1].$$

   We have: $\mathbf{E}[T_1] = 1/\lambda_A$.

We can find $\mathbf{E}[S_0]$ via the law of total expectations:

$$\begin{aligned}
\mathbf{E}[S_0] = \mathbf{E}[\min\{T_0, 1\}] &= \mathbf{P}(T_0 \leq 1)\mathbf{E}[T_0 \mid T_0 \leq 1] + \mathbf{P}(T_0 > 1)\mathbf{E}[1|T_0 > 1] \\
&= \left(1 - e^{-\lambda_A}\right) \int_0^1 t f_{T|T_0 \leq 1}(t)\, dt + e^{-\lambda_A} \\
&= \left(1 - e^{-\lambda_A}\right) \int_0^1 t \frac{\lambda_A e^{-\lambda_A t}}{(1 - e^{-\lambda_A})}\, dt + e^{-\lambda_A} \\
&= \int_0^1 t\lambda_A e^{-\lambda_A t}\, dt + e^{-\lambda_A} \\
&= \frac{1}{\lambda_A} \int_0^1 t\lambda_A^2 e^{-\lambda_A t}\, dt + e^{-\lambda_A} \\
&= \frac{1}{\lambda_A} \left(1 - e^{-\lambda_A} - \lambda_A e^{-\lambda_A}\right) + e^{-\lambda_A} \\
&= \frac{1}{\lambda_A} \left(1 - e^{-\lambda_A}\right)
\end{aligned}$$

where the above integral is evaluated by manipulating the integrand into an Erlang order 2 PDF and equating the integral of this PDF from 0 to 1 to the probability that there are 2 or more arrivals in the first hour (i.e. $\mathbf{P}(Y_2 < 1) = 1 - \mathbf{P}(0, 1) - \mathbf{P}(1, 1)$). Alternatively, one can integrate by parts and arrive at the same result.

Combining the above expectations:

$$\mathbf{E}[Q] = \mathbf{E}[S_0] + \mathbf{E}[T_1] = \frac{1}{\lambda_A}\left(1 - e^{-\lambda_A}\right) + \frac{1}{\lambda_A} = \frac{1}{\lambda_A}\left(2 - e^{-\lambda_A}\right).$$

(c) **(5 points)** What is the expected value of the time until each one of them has sent at least one email? (Note that we count time starting from time 0, and we take into account any emails possibly sent out by Alice during the interval $[0, 1]$.)

Solution: Define U as the time from $t = 0$ until each person has sent at least one email.

Define V as the remaining time from when Bob arrives (time 1) until each person has sent at least one email (so $V = U - 1$).

Define S as the time until Bob sends his first email after time 1.

Define the event $A = \{\text{Alice sends one or more emails in the time interval } [0, 1]\} = \{Y_1 \leq 1\}$, where Y_1 is the time Alice sends her first email.

Define the event $B = \{\text{After time 1, Bob sends his next email before Alice does}\}$, which is equivalent to the event where the next arrival in the merged process from Alice and Bob's orginal processes (starting from time 1) comes from Bob's process.

We have:

$$\mathbf{P}(A) = \mathbf{P}(Y_1 \leq 1) = 1 - e^{-\lambda_A}$$

$$\mathbf{P}(B) = \frac{\lambda_B}{\lambda_A + \lambda_B}.$$

Then,

$$\begin{aligned}
\mathbf{E}[U] &= \mathbf{P}(A)\mathbf{E}[U \mid A] + \mathbf{P}(A^c)\mathbf{E}[U \mid A^c] \\
&= (1-e^{-\lambda_A})(1+\mathbf{E}[V \mid A]) + e^{-\lambda_A}(1+\mathbf{E}[V \mid A^c]) \\
&= (1-e^{-\lambda_A})(1+\mathbf{E}[V \mid A]) + e^{-\lambda_A}(1+\mathbf{P}(B \mid A^c)\mathbf{E}[V \mid B \cap A^c] + \mathbf{P}(B^c \mid A^c)\mathbf{E}[V \mid B^c \cap A^c]) \\
&= (1-e^{-\lambda_A})(1+\mathbf{E}[V \mid A]) + e^{-\lambda_A}(1+\mathbf{P}(B)\mathbf{E}[V \mid B \cap A^c] + \mathbf{P}(B^c)\mathbf{E}[V \mid B^c \cap A^c]) \\
&= (1-e^{-\lambda_A})(1+\mathbf{E}[V \mid A]) + e^{-\lambda_A}\left(1+\frac{\lambda_B}{\lambda_A+\lambda_B}\mathbf{E}[V \mid B \cap A^c] + \frac{\lambda_A}{\lambda_A+\lambda_B}\mathbf{E}[V \mid B^c \cap A^c]\right).
\end{aligned}$$

Note that $\mathbf{E}[V \mid B^c \cap A^c]$ is the expected value of the time until each of them sends one email after time 1 (since, given A^c, Alice did not send any in the interval $[0,1]$) and given Alice sends an email before Bob. Then this is the expected time until an arrival in the merged process followed by the expected time until an arrival in Bob's process. So, $\mathbf{E}[V \mid B^c \cap A^c] = \frac{1}{\lambda_A+\lambda_B} + \frac{1}{\lambda_B}$.
Similarly, $\mathbf{E}[V \mid B \cap A^c]$ is the time until each sends an email after time 1, given Bob sends an email before Alice. So $\mathbf{E}[V \mid B \cap A^c] = \frac{1}{\lambda_A+\lambda_B} + \frac{1}{\lambda_A}$.
Also, $\mathbf{E}[V \mid A]$ is the expected time it takes for Bob to send his first email after time 1 (since, given A, Alice already sent an email in the interval $[0,1]$). So $\mathbf{E}[V \mid A] = \mathbf{E}[S] = 1/\lambda_B$. Combining all of this with the above, we have:

$$\begin{aligned}
\mathbf{E}[U] &= (1-e^{-\lambda_A})(1+1/\lambda_B) \\
&\quad + e^{-\lambda_A}\left(1+\frac{\lambda_B}{\lambda_A+\lambda_B}\left(\frac{1}{\lambda_A+\lambda_B}+\frac{1}{\lambda_A}\right) + \frac{\lambda_A}{\lambda_A+\lambda_B}\left(\frac{1}{\lambda_A+\lambda_B}+\frac{1}{\lambda_B}\right)\right).
\end{aligned}$$

(d) **(5 points)** Given that a total of 10 emails were sent during the interval $[0,2]$, what is the probability that exactly 4 of them were sent by Alice?
Solution:

$$\begin{aligned}
\mathbf{P}(\text{Alice sent 4 in } [0,2] \mid \text{total 10 sent in } [0,2]) &= \frac{\mathbf{P}(\text{Alice sent 4 in } [0,2] \cap \text{total 10 sent in } [0,2])}{\mathbf{P}(\text{total 10 sent in } [0,2])} \\
&= \frac{\mathbf{P}(\text{Alice sent 4 in } [0,2] \cap \text{Bob sent 6 } [0,2])}{\mathbf{P}(\text{total 10 sent in } [0,2])} \\
&= \frac{\left(\frac{(2\lambda_A)^4 e^{-2\lambda_A}}{4!}\right)\left(\frac{(\lambda_B)^6 e^{-\lambda_B}}{6!}\right)}{\frac{(2\lambda_A+\lambda_B)^{10} e^{-2\lambda_A+\lambda_B}}{10!}} \\
&= \binom{10}{4}\left(\frac{2\lambda_A}{2\lambda_A+\lambda_B}\right)^4\left(\frac{\lambda_B}{2\lambda_A+\lambda_B}\right)^6.
\end{aligned}$$

As the form of the solution suggests, the problem can be solved alternatively by computing the probability of a single email being sent by Alice, given it was sent in the interval $[0,2]$. This can be found by viewing the number of emails sent by Alice in $[0,2]$ as the number of arrivals arising from a Poisson process with twice the rate ($2\lambda_A$) in an interval of half the duration (particularly, the interval $[1,2]$), then merging this process with Bob's process. Then the probability that an email sent in the interval $[0,2]$ was sent by Alice is the probability that an arrival in this new merged process came from the newly constructed $2\lambda_A$ rate process:

$$p = \frac{2\lambda_A}{2\lambda_A + \lambda_B}.$$

Then, out of 10 emails, the probability that 4 came from Alice is simply a binomial probability with 4 successes in 10 trials, which agrees with the solution above.

5. **(5 points)** Suppose that $\lambda_A = 4$. Use Chebyshev's inequality to find an upper bound on the probability that Alice sent at least 5 emails during the time interval $[0, 1]$. Does the Markov inequality provide a better bound?

   Solution:

   Let N be the number of emails Alice sent in the interval $[0, 1]$. Since N is a Poisson random variable with parameter λ_A,

   $$\mathbf{E}[N] = \text{var}(N) = \lambda_A = 4.$$

   To apply the Chebyshev inequality, we recognize:

   $$\mathbf{P}(N \geq 5) = \mathbf{P}(N - 4 \geq 1) \leq \mathbf{P}(|N - 4| \geq 1) \leq \frac{\text{var}(N)}{1^2} = 4.$$

   In this case, the upper-bound of 4 found by application of the Chebyshev inequality is uninformative, as we already knew $\mathbf{P}(N \geq 5) \leq 1$.

   To find a better bound on this probability, use the Markov inequality, which gives:

   $$\mathbf{P}(N \geq 5) \leq \frac{\mathbf{E}[N]}{5} = \frac{4}{5}.$$

6. **(5 points)** You do not know λ_A but you watch Alice for an hour and see that she sent exactly 5 emails. Derive the maximum likelihood estimate of λ_A based on this information.

   Solution:

   $$\begin{aligned} \hat{\lambda}_A &= \arg\max_{\lambda} \log\left(p_N(5; \lambda)\right) \\ &= \arg\max_{\lambda} \log\left(\frac{\lambda^5 e^{-\lambda}}{5!}\right) \\ &= \arg\max_{\lambda} -\log(5!) + 5\log(\lambda) - \lambda. \end{aligned}$$

   Setting the first derivative to zero

   $$\frac{5}{\lambda} - 1 = 0$$

   $$\hat{\lambda}_A = 5.$$

7. **(5 points)** We have reasons to believe that λ_A is a large number. Let N be the number of emails sent during the interval $[0, 1]$. Justify why the CLT can be applied to N, and give a precise statement of the CLT in this case.

   Solution: With λ_A large, we assume $\lambda_A \gg 1$. For simplicity, assume λ_A is an integer. We can divide the interval $[0, 1]$ into λ_A disjoint intervals, each with duration $1/\lambda_A$, so that these intervals span the entire interval from $[0, 1]$. Let N_i be the number of arrivals in the ith such interval, so that the N_i's are independent, identically distributed Poisson random variables with parameter 1. Since N is defined as the number of arrivals in the interval $[0, 1]$, then $N = N_1 + \cdots + N_{\lambda_A}$. Since $\lambda_A \gg 1$, then N is the sum of a large number of independent and identically distributed random variables, where the distribution of N_i does not change as the number of terms in the sum increases. Hence, N is approximately normal with mean λ_A and variance λ_A.

   If λ_A is not an integer, the same argument holds, except that instead of having λ_A intervals, we have an integer number of intervals equal to the integer part of λ_A ($\bar{\lambda}_A$ =floor(λ_A)) of length $1/\lambda_A$ and an extra interval of a shorter length $(\lambda_A - \bar{\lambda}_A)/\lambda_A$.

   Now, N is a sum of λ_A independent, identically distributed Poisson random variables with parameter 1 added to another Poisson random variable (also independent of all the other Poisson random variables) with parameter $(\lambda_A - \bar{\lambda}_A)$. In this case, N would need a small correction to apply the central limit theorem as we are familiar with it; however, it turns out that even without this correction, adding the extra Poisson random variable does not preclude the distribution of N from being approximately normal, for large λ_A, and the central limit theorem still applies.

   To arrive at a precise statement of the CLT, we must "standardize" N by subtracting its mean then dividing by its standard deviation. After having done so, the CDF of the standardized version of N should converge to the standard normal CDF as the number of terms in the sum approaches infinity (as $\lambda_A \to \infty$).

   Therefore, the precise statement of the CLT when applied to N is:

$$\lim_{\lambda_A \to \infty} \mathbf{P}\left(\frac{N - \lambda_A}{\sqrt{\lambda_A}} \le z\right) = \Phi(z)$$

   where $\Phi(z)$ is the standard normal CDF.

8. **(5 points)** Under the same assumption as in last part, that λ_A is large, you can now pretend that N is a normal random variable. Suppose that you observe the value of N. Give an (approximately) 95% confidence interval for λ_A. State precisely what approximations you are making.

   Possibly useful facts: The cumulative normal distribution satisfies $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$.

   Solution: We begin by estimating λ_A with its ML estimator $\hat{\Lambda}_A = N$, where $\mathbf{E}[N] = \lambda_A$. With λ_A large, the CLT applies, and we can assume N has an approximately normal distribution. Since $\text{var}(N) = \lambda_A$, we can also approximate the variance of N with ML estimator for λ_A, so $\text{var}(N) \approx N$, and $\sigma_N \approx \sqrt{N}$.

To find the 95% confidence interval, we find β such that:

$$\begin{aligned} 0.95 &= \mathbf{P}\left(|N - \lambda_A| \le \beta\right) \\ &= \mathbf{P}\left(\frac{|N - \lambda_A|}{\sqrt{N}} \le \frac{\beta}{\sqrt{N}}\right) \\ &\approx 2\Phi\left(\frac{\beta}{\sqrt{N}}\right). \end{aligned}$$

So, we find:

$$\beta \approx \sqrt{N}\Phi^{-1}(0.975) = 1.96\sqrt{N}.$$

Thus, we can write:

$$\mathbf{P}(N - 1.96\sqrt{N} \le \lambda_A \le N + 1.96\sqrt{N}) \approx 0.95.$$

So, the approximate 95% confidence interval is: $[N - 1.96\sqrt{N}, N + 1.96\sqrt{N}]$.

9. You are now told that λ_A is actually the realized value of an exponential random variable Λ, with parameter 2:

$$f_\Lambda(\lambda) = 2e^{-2\lambda}, \qquad \lambda \ge 0.$$

(a) **(5 points)** Find $\mathbf{E}[N^2]$.

Solution:

$$\begin{aligned} \mathbf{E}[N^2] = \mathbf{E}[\mathbf{E}[N^2 \mid \Lambda]] &= \mathbf{E}[\mathrm{var}(N \mid \Lambda) + (\mathbf{E}[N \mid \Lambda])^2] \\ &= \mathbf{E}[\Lambda + \Lambda^2] \\ &= \mathbf{E}[\Lambda] + \mathrm{var}(\Lambda) + (\mathbf{E}[\Lambda])^2 \\ &= \frac{1}{2} + \frac{2}{2^2} \\ &= 1. \end{aligned}$$

(b) **(5 points)** Find the linear least squares estimator of Λ given N.

Solution:

$$\hat{\Lambda}_{\mathrm{LLMS}} = \mathbf{E}[\Lambda] + \frac{\mathrm{cov}(N, \Lambda)}{\mathrm{var}(N)}(N - \mathbf{E}[N]).$$

Solving for the above quantities:

$$\begin{aligned} \mathbf{E}[\Lambda] &= \frac{1}{2} \\ \mathbf{E}[N] &= \mathbf{E}[\mathbf{E}[N \mid \Lambda]] = \mathbf{E}[\Lambda] = \frac{1}{2}. \\ \mathrm{var}(N) &= \mathbf{E}[N^2] - (\mathbf{E}[N])^2 = 1 - \frac{1}{2^2} = \frac{3}{4}. \\ \mathrm{cov}(N, \Lambda) &= \mathbf{E}[N\Lambda] - \mathbf{E}[N]\mathbf{E}[\Lambda] = \mathbf{E}[\mathbf{E}[N\Lambda \mid \Lambda]] - (\mathbf{E}[\Lambda])^2 = \mathbf{E}[\Lambda^2] - (\mathbf{E}[\Lambda])^2 = \mathrm{var}(\Lambda) = \frac{1}{4}. \end{aligned}$$

Substituting these into the equation above:

$$\begin{aligned}
\hat{\Lambda}_{\text{LLMS}} &= \mathbf{E}[\Lambda] + \frac{\text{cov}(N,\Lambda)}{\text{var}(N)}(N - \mathbf{E}[N]) \\
&= \frac{1}{2} + \frac{1/4}{3/4}\left(N - \frac{1}{2}\right) \\
&= \frac{1}{3}\left(N + 1\right).
\end{aligned}$$

Final Exam Statistics