CS2810 Day 26

Admin:
TRACE @ ~50%, please do this!

Review:
Normal Distribution
- Central Limit Theorem
- CDF

Hypothesis Testing
- Whats a p-value?
- Experimental Bias
- T-Tests
    - one vs two sided
- Chi Square Test
- Multiple Comparison Correction

Covariance
- Covariance Matrix
- Correlation
- Independence & Correlation / Covariance

Bayes Rule Problems:
- Binary Variables
    - p(covid | test positive for covid)?
- Parameterized Likelihoods
    - poisson: traffic flow rate problem on HW
    - binomial: coconut special ICA (day 23)

Bayes Nets:
- Identifying (conditional) independence
- How to compute conditional prob
    - step 1: rewrite without conditional
    - step 2c: build joint distribution table
    - step 3c: marginalize joint probs from step 1 via joint distribution table

Problem 1: "Five Second Rule"
Normal Distribution / Central Limit Theorem / CDF

Assume the snacks my daughter drops on the floor which the dog then eats is a poisson distribution with lambda = 15 snacks / day. (This model has one big glaring assumption problem ... what is it?)

Estimate the probability that the dog eats more than 7 pounds of dropped food over a whole year.

Write out and evaluate any assumptions you deem necessary.

Problem 2: Hypothesis Testing

Two artists sell their work at auction.

Artist A's works go for (in thousands of dollars):

3, 4, 5

Artist B's works go for (in thousands of dollars):

5, 6, 7, 8

- Identify 3 sources of experimental bias under which this data could've been collected
    - location art was sold
    - sequencing of sales
    - preferences of buyers
- Perform any hypothesis test you deem relevant to answer the question: does B's work sell for more than A's?
    - whats is a type 1/2 errors in this case? can we say anything about prob of type 1/2 errors?

$$T = \frac{\bar{A} - \bar{B}}{S}$$

$$S^2 = \frac{\hat{\sigma}_A^2}{N_A} + \frac{\sigma_B^2}{N_B} = \sqrt{.75}$$

$$T = \frac{4 - 6.5}{\sqrt{3/4}} = -2.88$$

does B's work sell for more than A's?

H0: \mu_b <= \mu_a
H1: ~~B's work sells for~~ more than A's (\mu_b > \mu_a)

reject H0, claim H1 is true
do not reject H0, (no claims)

$A = 3 \quad 4 \quad 5$

$$\hat{\sigma}_A^2 = \frac{1}{N-1} \sum_i (A_i - \bar{A})^2$$

$$= \frac{1}{2}(1^2 + 0^2 + 1^2) = 1$$

$$\hat{\sigma}_B^2 = \frac{1}{N-1}\sum\left(B_i - \bar{B}\right)^2$$

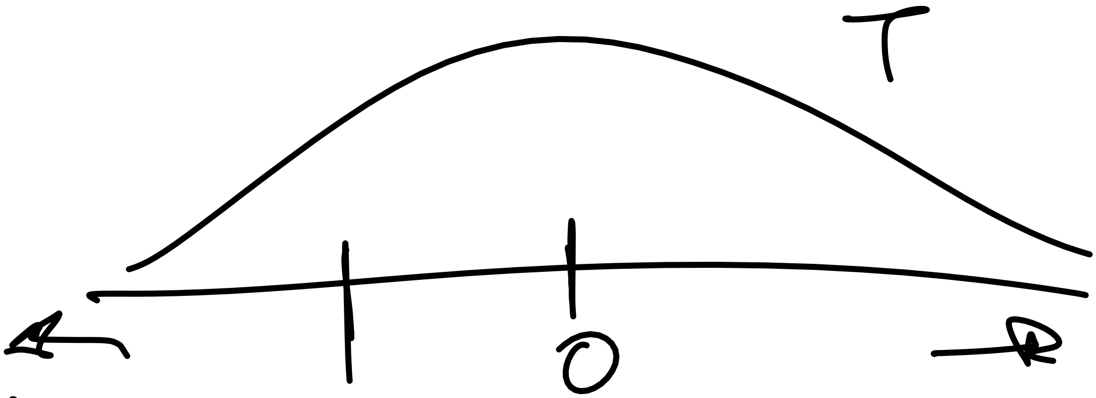
$$= \frac{1}{3}\left[(5-6.5)^2 + (6-6.5)^2 + (7-6.5)^2 + (8-6.5)^2\right] = \frac{2}{3}\left(1.5^2 + .5^2\right)$$

5 6 7 8

$DF = N_A + N_B - 2$

$= 3 + 4 - 2$

$= 5$



$\mu_A < \mu_b$

LESS CONSISTENT

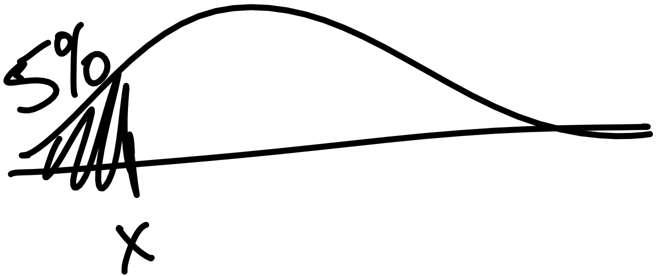
H_0

$\mu_A > \mu_B$

MORE CONSISTENT

CDF(.05)

Df	4	5	6	7
X	-1.7	-1.6	-1.5	1.4



$P_{VAL} = .018 \quad \alpha = .05$

REJECT H_0, CLAIM H_1

$\mu_B > \mu_A$

Problem 3: Check, please!

The covariance (left) correlation (right) and mean (bottom left) of features describing diner bills is given below.

- Explain why tip (the total amount tipped) has a positive correlation with total_bill while tip_perc (tip as a percent of total) has a negative correlation.
- Describe the total_bill, smoker and size of the dining party which who gives the lowest tip (absolute, not perc)
- (+) The value 4 is most likely to belong to which of the five features below? Explain

	total_bill	tip	smoker	size	tip_perc
total_bill	79.252939	8.323502	0.371388	5.065983	-0.184107
tip	8.323502	1.914455	0.003992	0.643906	0.028931
smoker	0.371388	0.003992	0.236845	-0.061644	0.000916
size	5.065983	0.643906	-0.061644	0.904591	-0.008298
tip_perc	-0.184107	0.028931	0.000916	-0.008298	0.003730

	total_bill	tip	smoker	size	tip_perc
total_bill	1.000000	0.675734	0.085721	0.598315	-0.338624
tip	0.675734	1.000000	0.005929	0.489299	0.342370
smoker	0.085721	0.005929	1.000000	-0.133178	0.030820
size	0.598315	0.489299	-0.133178	1.000000	-0.142860
tip_perc	-0.338624	0.342370	0.030820	-0.142860	1.000000

mean of each feat:

total_bill	19.785943
tip	2.998279
smoker	0.381148
size	2.569672
tip_perc	0.160803

Problem 4: Sample mean / cov / corr compute

Compute the unbiased sample mean, covariance matrix and correlation matrices for the observations below

$x = 4, 7, 9, 41$
$y = 3, 2, 1, 0$

(each column above is a pair of observations $(x0, y0) = (4,3)$, $(x1, y1) = (7, 2)$, ...

Problem 5: Bayes

$P(A)$ $A \longrightarrow X$

Aliens, were they to exist on mars, would show up in .001 of photographs taken of the martian surface.

$P(X|A)$

In the event Aliens don't exist, they'd never appear.

If we've taken 57 pictures of the surface of mars and an Alien hasn't shown up in any, whats the probability they exist?

Make any assumptions (i.e. a prior probability for aliens) you deem necessary.
(Its a big drawback to bayesian analysis that we need to make a prior distribution ... feels rather subjective to estimate like this, right?)

$P(A=1) = 80\%$

A=1 ALIENS EXIST ON MARS

X= # PHOTOS W/ ALIENS IN 57

$P(X|A=1) = \text{BINOM}(n=57, p=.001)$

$P(X|A=0) = \text{BINOM}(n=57, p=0)$

$$P(A=1 \mid X=0) = \frac{P(X=0 \mid A=1)\,P(A=1)}{P(X=0)}$$

$$= \frac{.944 \cdot .8}{.944 \cdot .8 + 1 \cdot .2} \approx .79$$

$$P(X=0 \mid A=1) = \text{Binom.PMF}(x=0, n=57, p=.001) = .944$$

$$P(X=0 \mid A=0) = 1$$

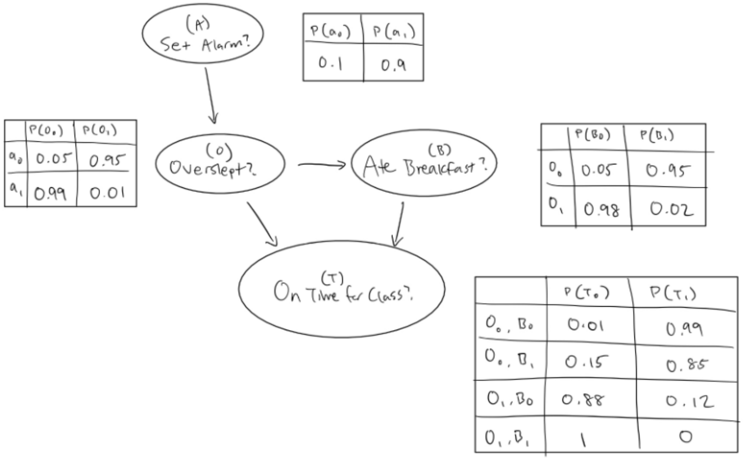
No pics aliens
given no aliens

$$P(X=0) = P(X=0\ A=0) + P(X=0\ A=1)$$

$$= P(X=0 \mid A=0)\,P(A=0) + P(X=0 \mid A=1)\,P(A=1)$$

Problem 6: Bayes Net

1. Compute the joint distribution table for the Bayes Net
2. Compute prob one is on time for class.
3. Compute prob one is on time for class given they didnt set their alarm.
4. Compute prob one is on time for class given they didnt set their alarm and skipped breakfast.
5. Can you explain, via intuition informed by the network to the right, how prob from questions 2/3 and questions 3/4 compare? (e.g. why is 3 higher / lower than 2?)



Bayes Net Credit:
li.mingle@northeastern.edu,
panos.a@northeastern.edu,
leonard.l@northeastern.edu,
hernandez.die@northeastern.edu

Problem 6: Bayes Net

1. Compute the joint distribution table for the Bayes Net
2. Compute prob one is on time for class.
3. Compute prob one is on time for class given they didnt set their alarm.
4. Compute prob one is on time for class given they didnt set their alarm and skipped breakfast.
5. Can you explain, via intuition informed by the network to the right, how prob from questions 2/3 and questions compare? (e.g. why is 3 higher / lower than 2?)

Bayes Net Credit:
li.mingle@northeastern.edu,
panos.a@northeastern.edu,
leonard.l@northeastern.edu,
hernandez.die@northeastern.edu

(T) On Time For Class?	(B) Ate Breakfast?	(O) Overslept?	(A) Set Alarm?	P(TBOA)
t0	b0	o0	a0	0.0000025
t0	b0	o0	a1	0.0004455
t0	b0	o1	a0	0.081928
t0	b0	o1	a1	0.0077616
t0	b1	o0	a0	0.0007125
t0	b1	o0	a1	0.1269675
t0	b1	o1	a0	0.0019
t0	b1	o1	a1	0.00018
t1	b0	o0	a0	0.0002475
t1	b0	o0	a1	0.0441045
t1	b0	o1	a0	0.011172
t1	b0	o1	a1	0.0010584
t1	b1	o0	a0	0.0040375
t1	b1	o0	a1	0.7194825
t1	b1	o1	a0	0
t1	b1	o1	a1	0

$$P(t_1 \mid a_0 b_0) = \frac{P(t_1 a_0 b_0)}{P(a_0 b_0)}$$