CS2810 Day 15

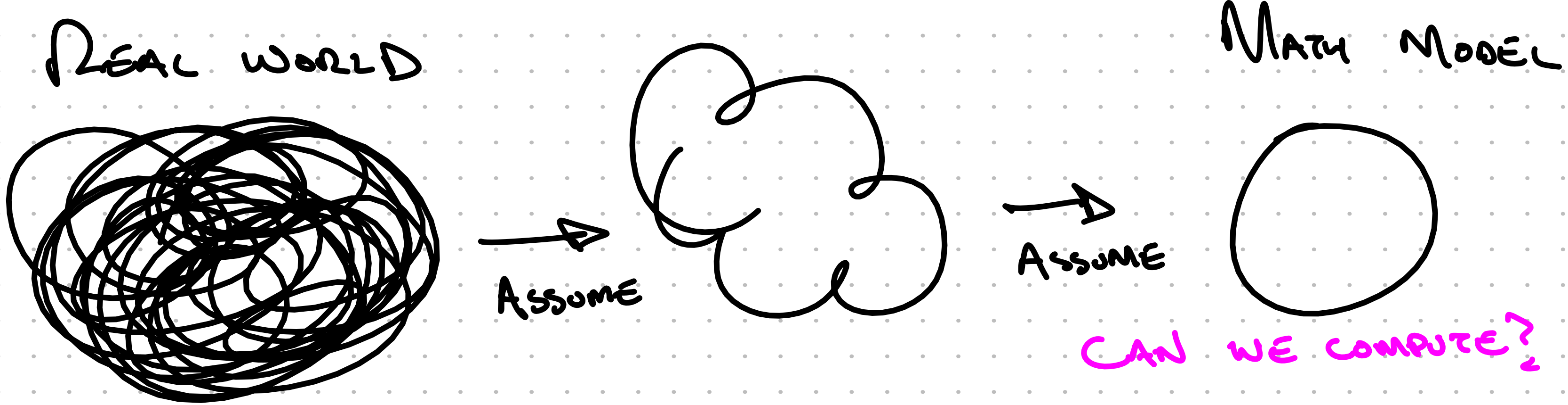
Admin:
HW6 released today (due date Mar 22 @ 11:59 PM, no late days accepted)
prob / stats calculator has new estimators / bias material from today
thank you for wearing masks

Content:
Binomial / Poisson Assumptions (HW6 practice)

Observations vs Ground Truth

Estimators
- What is a biased estimator?

Bessel's correction
- an unbiased way to estimate variance



CAN WE COMPUTE?

DO WE TRUST SIMPLE MODEL?

Today's skills:
- interpret / evaluate model assumptions in the context of a problem
- estimate model parameters
    - Poisson:
        lambda: the "rate" at which events occur
    - Binomial
        n: number of trials
        p: probability of success of each trial

Modeling with Binomial Distribution

Binomial:
How many "successes" occur in N total binary trials?

Parameters:
n=number of trials
p=prob of each trial's success

Assumes:
1. output of each trial is binary
2. outcome of every two trials is independent
3. each trial is "identically distributed"
    - has the same prob of success

Application:
A basketball player will take 100 free throws next season, how many do they make?

Model:
Binomial
    n=100
    p=average free throws from this season

Assumption violations:

2. emotions high / low after a make / miss impacts chance of next shot (independence)

3. hard foul will change prob of free throw (but not all fouls are hard)

3. context of game change prob success

Modelling with Poisson

Poisson:
How many events occur in a given time window?

Parameters:
lambda = "rate" of how many events typically occur in the time window

Assumes:
1. Occurance of one event doesn't impact occurance of future events
2. Rate that events are expected to occur is constant

Application:
How many groups will walk into store in any hour they're open during the week?

Model:
lambda = average of how many customers walked into store per hour during the previous year

Assumption violations:

2. there are busier times for the store than others

1. if store is busy more / less people might come

The Poisson Rate Parameter scales linearly

If

the number of car accidents a day in all of Boston is poisson distributed with lambda = 24

then....

the number of car accidents an hour in all of Boston is Poisson distributed with lambda = 1

ICA 1

The following table gives the number of groups entering a store each day:

Mon: 32   Tues: 40   Weds: 20   Thurs: 42   Fri: 41   Sat: 102   Sun 103

1. Assuming the store is open 8 hours a day, build a Poisson Distribution (estimate lambda) over the number of groups entering each hour.

2. Assuming your model, compute the probability that exactly 20 groups enter the store in an hour

3. Does the data above seem consistent with the Poisson assumptions?
-Rate that events are expected to occur is constant
-Occurance of one event doesn't impact occurance of future events

PROB K EVENTS

WHEN RATE IS λ

$$\frac{\lambda^k e^{-\lambda}}{k!}$$

Mon: 32 Tues: 40 Weds: 20 Thurs: 42 Fri: 41 Sat: 102 Sun 103

$$32+40+20+42+41+102+103 = \frac{380 \text{ GROUPS}}{7.8 \text{ HOURS}} \approx 6.78 \text{ GROUPS/HOUR}$$

③ PROB 20 GROUPS ENTER IN HR:

$$\frac{\lambda^{k} e^{-\lambda}}{k!} = \frac{6.78^{20} e^{-6.78}}{20!} \approx \text{JUST ABOVE } 0$$

Think alone:

You observe some fish in a new pond:

- one 3 pound fish
- one 5 pound fish

Do you, with certainty, know the expected value of the weights of fish in the pond you'd catch?

Observations vs Ground Truth

Observed data:
Collected in an experiment

Ground Truth data:
Describes the precise, absolutely true state
- rarely known

FISH WEIGHT

$x_1 = 3 \quad x_2 = 7 \quad x_3 = 5$

LET F BE RANDOM VARIABLE FISH WEIGHT IN POND

$E[F] = ?$

Estimators

An Estimator is a function of observations which outputs an estimate of some ground truth variable.

FISH WEIGHT

$f_1 = 3 \quad f_2 = 7 \quad f_3 = 5$

LET F BE RANDOM VARIABLE FISH WEIGHT IN POND

$$\frac{f_1 + f_2 + f_3}{N} = \frac{3+7+5}{3} = 5$$

$$E[F] = ?$$

"SAMPLE MEAN"

$$\bar{X} = \frac{1}{N} \sum_i x_i$$ IS AN ESTIMATOR FOR $E[x]$

i-TH OBSERVATION

TOTAL OBSERVATIONS

EACH OBSERVATION GETS LOWERCASE w/ INDEX

$$E[x] = \sum_i x_i P(x_i)$$

SUM OVER OUTCOME

⚠ SAME NOTATION FOR OUTCOMES OF EXPERIMENT

LIGHTENING ICA (IN YOUR HEAD)

YOU OBSERVE FISH WEIGHTS

$X_1 = 10 \quad X_2 = 20 \quad X_3 = 30$

→ ESTIMATE $E[x]$ w/ SAMPLE MEAN

→ ARE YOU CERTAIN THIS ESTIMATE IS EXACTLY EQUAL TO $E[x]$?

ESTIMATING EXPECTED VALUE

SUPPOSE YOU OBSERVE FISH WEIGHTS

$x_1 = 3 \qquad x_2 = 5 \qquad x_3 = 4$

Xi IS OUTCOME

$$\bar{x} = \frac{1}{N} \sum_i x_i \xrightarrow{\text{ESTIMATES}} E[x] = \sum_i x_i P(x_i)$$

SAMPLE MEAN

Xi IS OBSERVATION

ESTIMATING VARIANCE

SUPPOSE YOU OBSERVE FISH WEIGHTS

$X_1 = 3 \qquad X_2 = 5 \qquad X_3 = 4$

$$\hat{\sigma}^2_{\text{BIAS}} = \frac{1}{N} \sum_i \left(X_i - \bar{X} \right)^2 \xrightarrow{\text{ESTIMATES}} \sigma^2 = \text{VAR}(x) = \sum_i \left(x_i - E[x] \right)^2 P(x_i)$$

ICA 2

COMPUTE SAMPLE MEAN + "SAMPLE VARIANCE" OF $N=5$ SIX-SIDED DIE ROLLS

$x_0 = 6 \quad x_1 = 3 \quad x_2 = 5 \quad x_3 = 4 \quad x_4 = 5$

$$\hat{\sigma}^2_{\text{BIAS}} = \frac{1}{N} \sum_i \left(x_i - \bar{x} \right)^2$$

$x_0 = 6 \quad x_1 = 3 \quad x_2 = 5 \quad x_3 = 4 \quad x_4 = 5$

$$\bar{X} = \frac{1}{N} \sum_i x_i = \frac{6+3+5+4+5}{5} = \frac{23}{5} = 4.6$$

$$\hat{\sigma}^2_{\text{BIAS}} = \frac{1}{N} \sum_i (x_i - \bar{x})^2 = \frac{1}{5}\left((6-4.6)^2 + (3-4.6)^2 + (5-4.6)^2 + (4-4.6)^2 + (5-4.6)^2 \right) = 1.04$$

REMEMBER:

IF X IS FAIR SIX SIDED DIE ROLL

$$E[X] = \sum_i x_i P(x_i) = \tfrac{1}{6}\cdot 1 + \tfrac{1}{6}\cdot 2 + \tfrac{1}{6}\cdot 3 + \tfrac{1}{6}\cdot 4 + \tfrac{1}{6}\cdot 5 + \tfrac{1}{6}\cdot 6$$

$$= 3.5$$

$$E[X^2] = \sum_i x_i^2 P(x_i) = \tfrac{1}{6}\cdot 1^2 + \tfrac{1}{6}\cdot 2^2 + \tfrac{1}{6}\cdot 3^2 + \tfrac{1}{6}\cdot 4^2 + \tfrac{1}{6}\cdot 5^2 + \tfrac{1}{6}\cdot 6^2$$

$$= \tfrac{1}{6}\ \ 1 + 4 + 9 + 16 + 25 + 36$$

$$= 91/6$$

$$\mathrm{VAR}(X) = E[X^2] - E[X]^2 = \frac{91}{6} - 3.5^2 = 2.9167$$

Bessel's Motivation

(Python)

UNBIASED ESTIMATORS

An estimator is unbiased if its expected value equals the ground truth target.

Is the sample mean an unbiased estimator? ... yes, let's prove it:

$$E[\bar{x}] = E\left[\frac{1}{N}\sum_i x_i\right] = \frac{1}{N}\left(E[x_0] + E[x_1] + \dots + E[x_{N-1}]\right)$$

$$= \frac{1}{N}\left(E[x] + E[x] + \dots + E[x]\right)$$

$$= E[x]$$

An unbiased estimator of variance (Bessel's Correction) USE N-1 INSTEAD OF N

Claim:

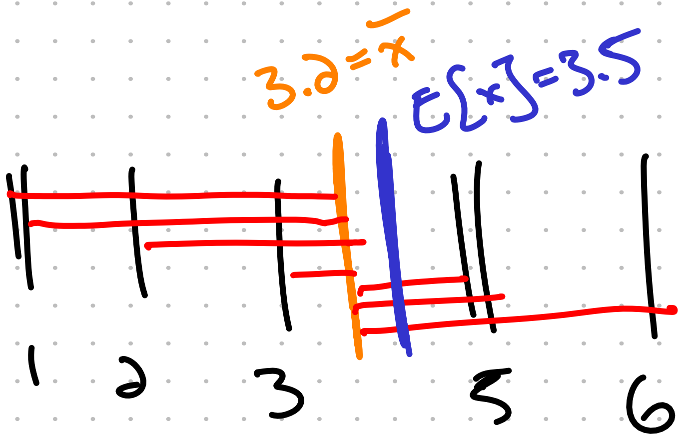
$$\hat{\sigma}^2_{BIAS} = \frac{1}{N}\sum_i (x_i - \bar{x})^2 \text{ IS BIASED (TOO SMALL)}$$

$$\hat{\sigma}^2_{BESSEL} = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 \text{ IS UNBIASED}$$

(WE WON'T PROVE IT...
BUT LET'S TEST IN PYTHON)

Bessel's Correction: Motivation

Why is $\hat{\sigma}^2_{\text{BIAS}} = \frac{1}{N}\sum_i \left(X_i - \bar{X}\right)^2$ often smaller than $\text{VAR}(X)$?



$\bar{X}$ is as close as possible to all X_i

$\bar{X}$ minimizes $\sum_i \left(x_i - \bar{x}\right)^2$

Bessel Motivation 2:

If we have a single observation, what can we say about variance?

$$\hat{\sigma}^2_{BIAS} = \frac{1}{N} \sum_i (x_i - \bar{x})^2 = \frac{1}{1} \cdot (4-4)^2 = 0$$

ICA 3: One more trip to the pond

The following are weights, in pounds, of fish you observe in a pond:

3, 5, 7, 1, 9, 8, 2

Let X be a Random Variable representing the weight of a fish in this pond

1. Give an unbiased estimate of E[x]

2. Give an unbiased estimate of Var(x)

3. Suppose a fish pops his head above the surface and claims, "Our average weight down here is 6 pounds". Incorporate his information into your unbisaed estimate of Var(x)

3 5 7 1 9 8 2

① $\bar{x} = \frac{1}{N} \sum_{1} x_i = \frac{3+5+7+1+9+8+2}{7} = 5$

② $\hat{\sigma}^2_{BESSEL} = \frac{1}{N-1} \sum_{i} (x_i - \bar{x})^2$

$= \frac{1}{7-1} \left((3-5)^2 + (5-5)^2 + (7-5)^2 + (1-5)^2 + (9-5)^2 + (8-5)^2 + (2-5)^2 \right)$

$= 9\,2/3$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i \left(x_i - E[x]\right)^2$$

$$=$$