Introduction to Statistics and Bayesian Networks

Soma Sekhar Dhavala
Iowa State University

what is a statistic

- any function of the data | mean, max, absolute value ...

why statistics

- to quantify uncertainty
- express randomness
- make an informed decision (however bad it may be :)]

(how likely is it that it rains today given that it rained yesterday and monsoon hit the western ghats)

Key terminology:

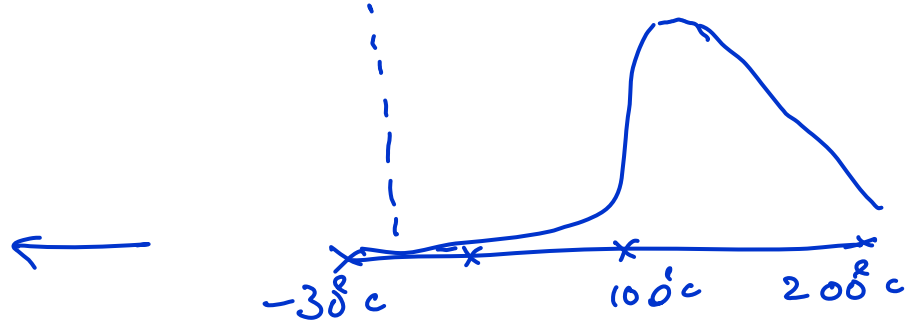
random variable: . the observations of which over a period of time are possibly different characterized by a distribution

distribution: a function of the random-variable, when evaluated at a possible outcome / realization of a r.v's, tells us how probable that r.v can be observed over a sufficient length of interval

e.g of a raw boiler temperature : (θ)

$-30^\circ C$ to $200^\circ C$

$f(\theta)$



room temparature

eg.

$f(99^\circ \le \theta \le 102) = 0.5$

$f(-30^\circ \le \theta \le -10^\circ c) = 0.0001$

$$\int_{\theta=-30^\circ c}^{200^\circ c} f(\theta)\, d\theta = 1 \qquad f(\theta) \geq 0 \quad \forall\, \theta$$

laws of probability

discrete random variables

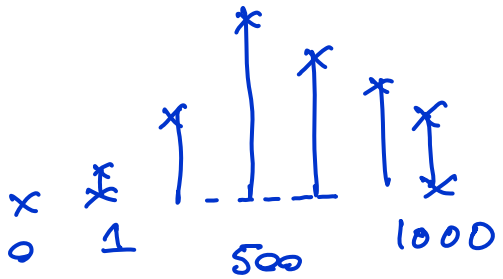
ATM transactions 0 to 1000

the distribution is now called

probability mass function

or

(pmf)



If the r.v is continous, it is prob. density function

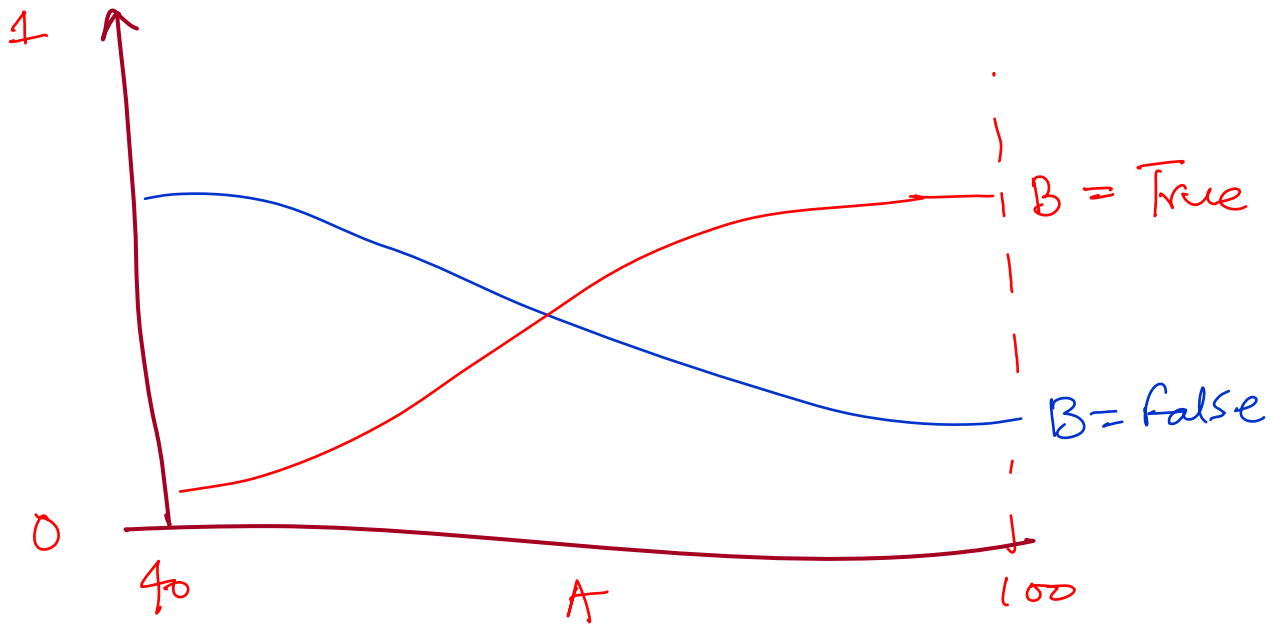
(pdf)

$P(\#ATM = 100) = 0.01$

Several r.v's (continous or discrete)

A : price of petrol/liter (Rs40/- to 100/-)

B : war on terrorism (True or false)



$P(A, B)$ is the joint distribution

If $P(A \mid B)$ is known to us, we can draw several inferences such as

Given that oil price is Rs 90/- per liter, which is more likely war on terrorism true or false

— x —

$P(B / A = 90\text{Rs})$??

$P(B = T / A = 90) > P(B = F / A = 90)$

what is a statistical model??

eg. Yield of Corn (is the variable of interest)
in tons/sq. mile

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon$$

- Y: observed quantity
- x_1: amount of nitrogen
- x_2: % rain fall
- ϵ: error terms
- x_1, x_2: control variables
- $\beta_1, \beta_2 \ldots$: are unknowns

Given the observations, what are the likely values of β's which could have possibly produced the observations that we have at hand

Different estimation techniques

- Maximum likelihood
- Maximum aposteriori
- Bayesian methods

Then we can optimize the parameters of the model to yield maximum gains

max Y w.r.t $(x_1, x_2 \cdots x_N)$

given β's

100 ppm Sucfar ~~with~~ under 10 mm rainfall will yield maximum returns

But how do we generate data

Design of experiments

fix the inputs (controlled / uncontrolled)

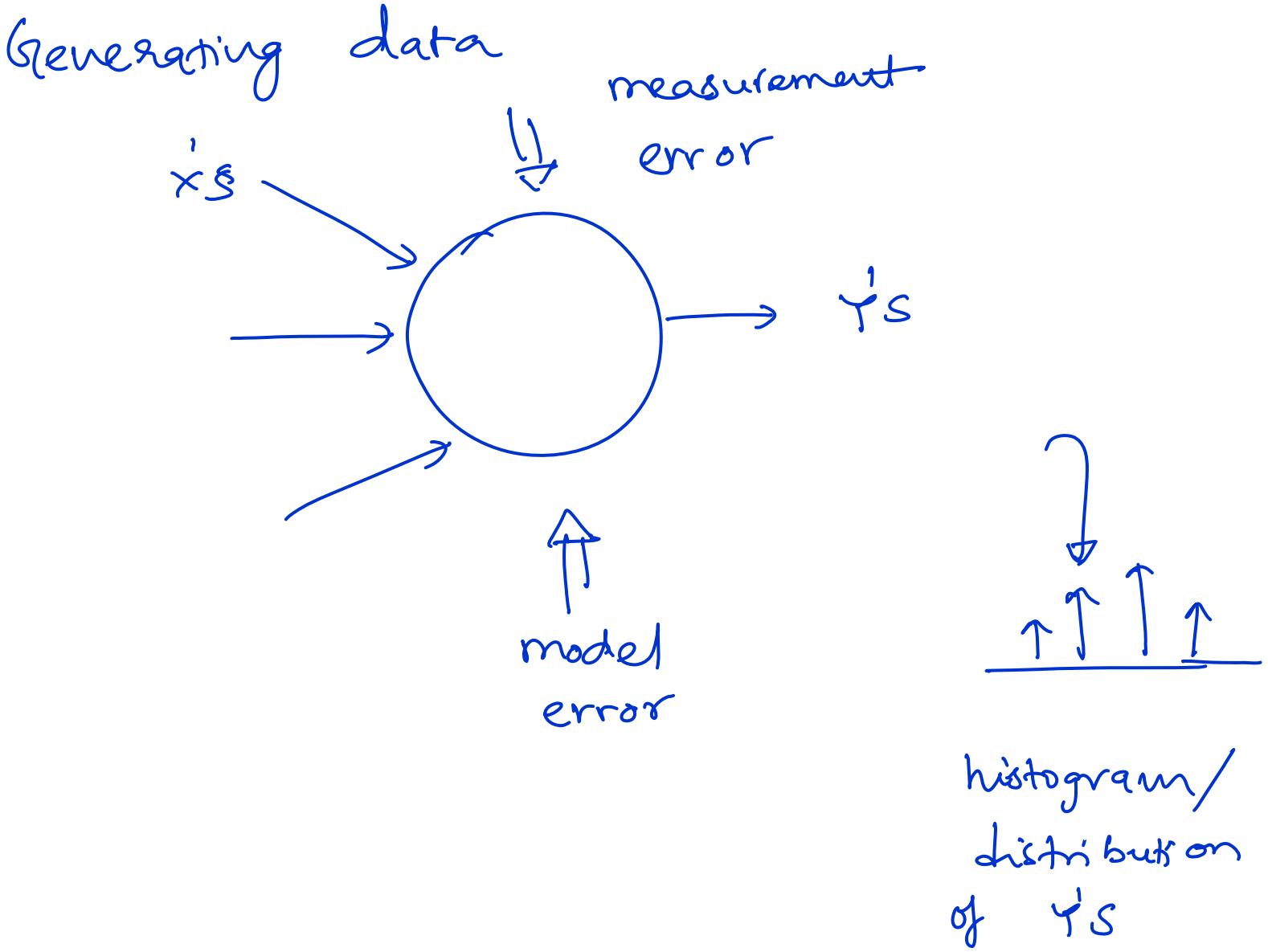
Set the objective of the experiment

Study the effects of varying the input on the outcome in a methodological fashion

$$f(\text{inputs}) \longrightarrow \text{output}$$

how many inputs

how many levels / factors each input takes

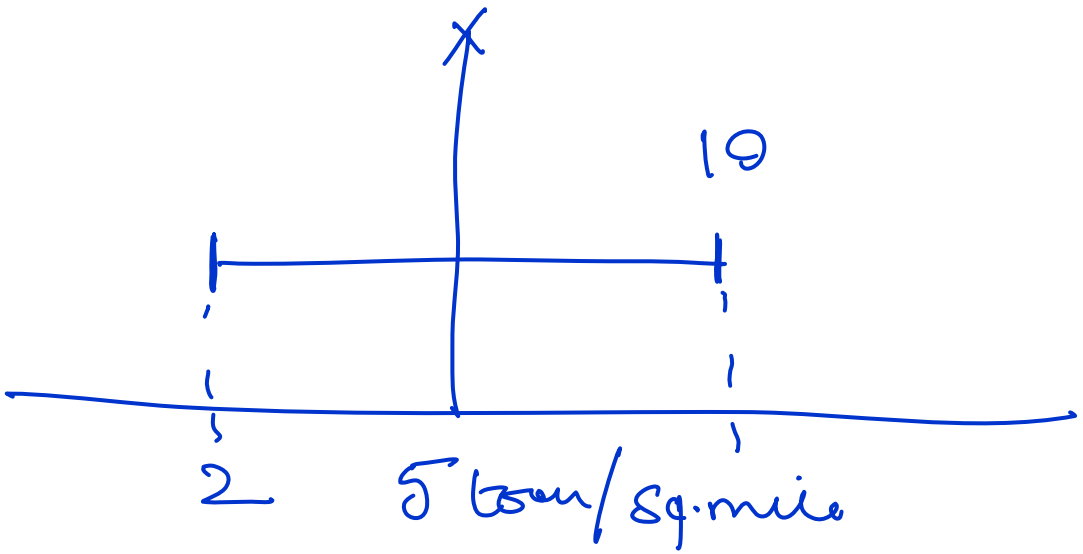

Generating data
measurement error
x's
y's
model error
histogram/ distribution of y's

Some inferences on the yield

→ avg. yield $\underbrace{E(Y)}$

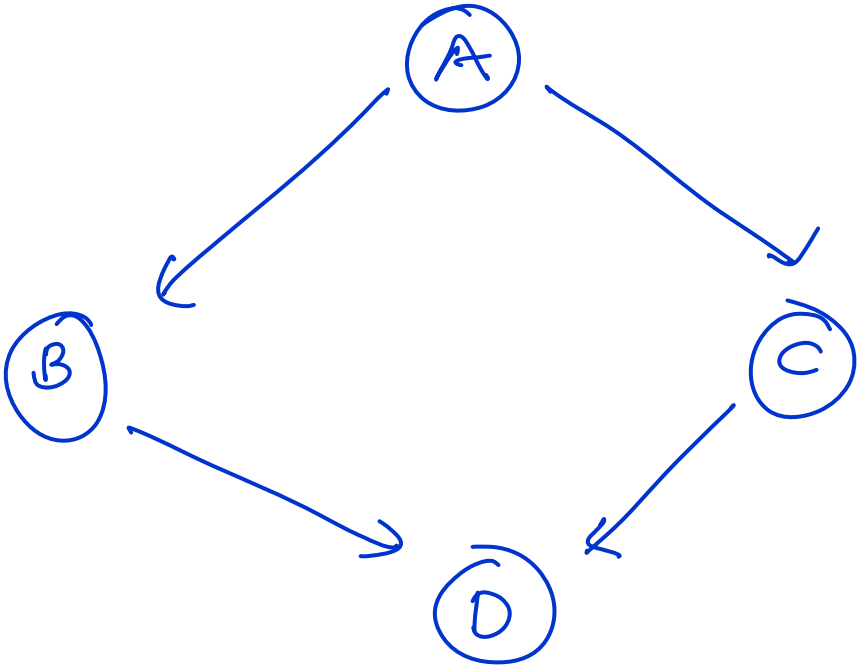
and

what is the confidence interval



95% confidence level

Bayes Networks



→ Parent - child relationship

→ nodes are random variables

→ Cause — effect analysis possible

either ways