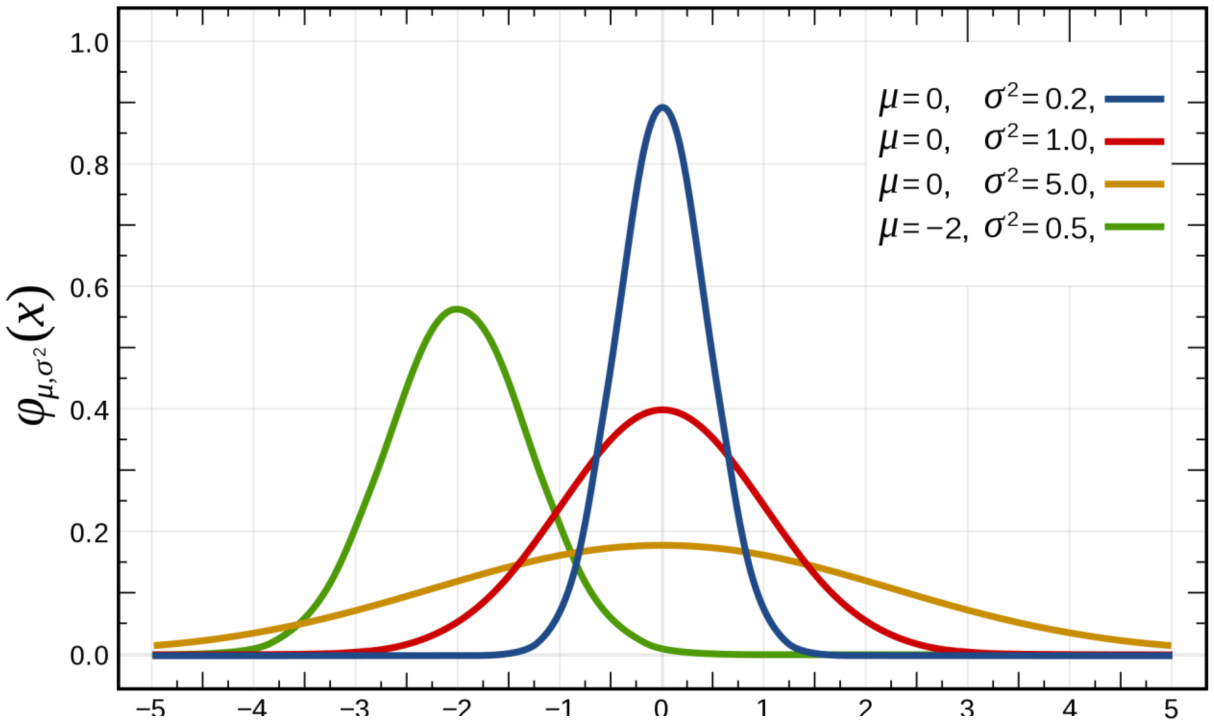
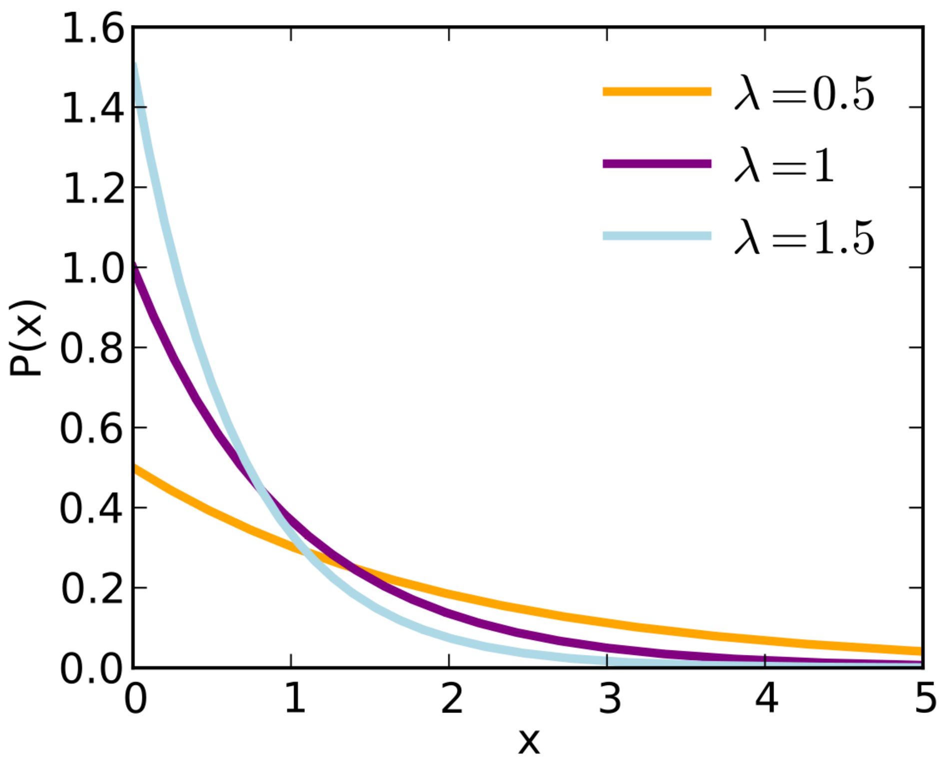
A Primer to Parameter Estimation

Srijith P K

Sampling and Estimation

- What you have learnt : X is a random variable following a distribution
  - what does this mean ?
- Given the parameter value of a distribution how the density/distribution function look.
- Sampling : If X follows a distribution how can we obtain different value X will take
- Parameter estimation : Given different values X take, how can we obtain the parameters of the underlying distribution

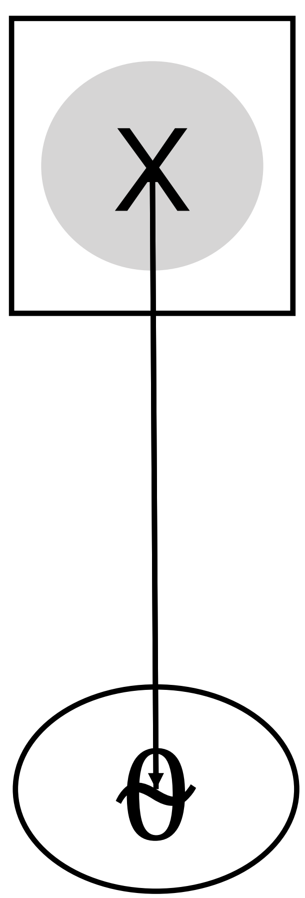


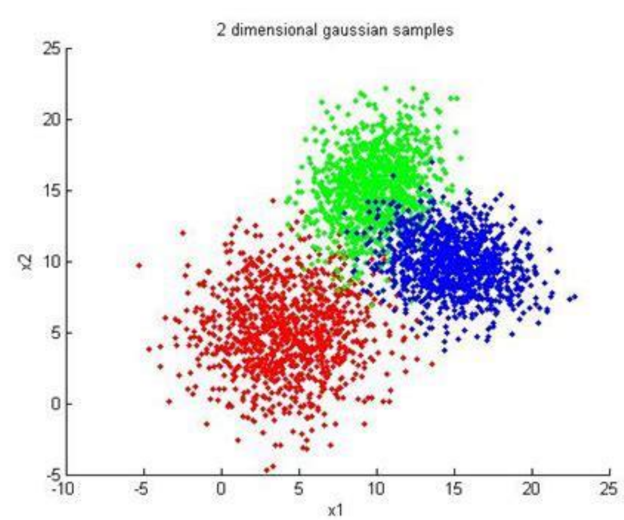


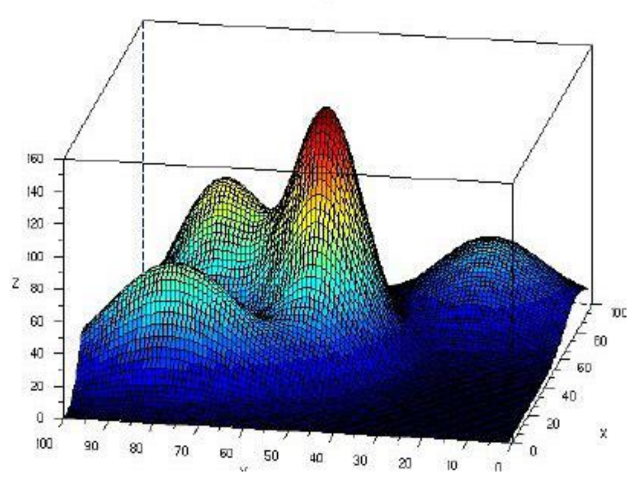
Sampling and Estimation

- Given data Statistics/ML aim to learn parameters of the underlying distribution from data
  - Gaussian mixture model, probabilistic graphical models, linear regression, logistic regression etc.
- Sampling is essential in probabilistic machine learning and Bayesian statistics
  - Latent dirichlet allocation, Gaussian process, probabilistic graphical model
  - Also used in computational physics, biology and many engineering disciplines.

    MCMC was placed in the top 10 most important algorithms of the 20th century

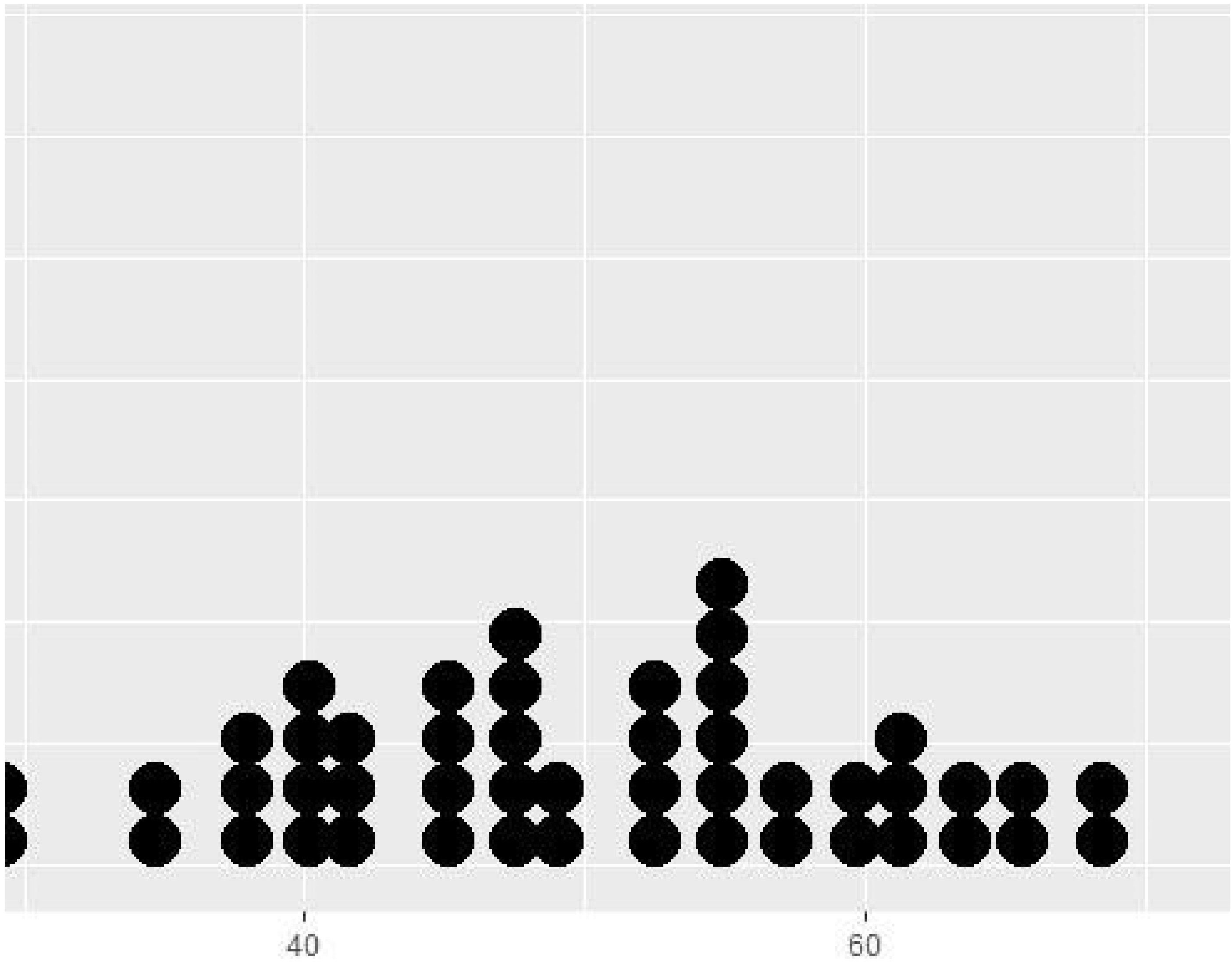






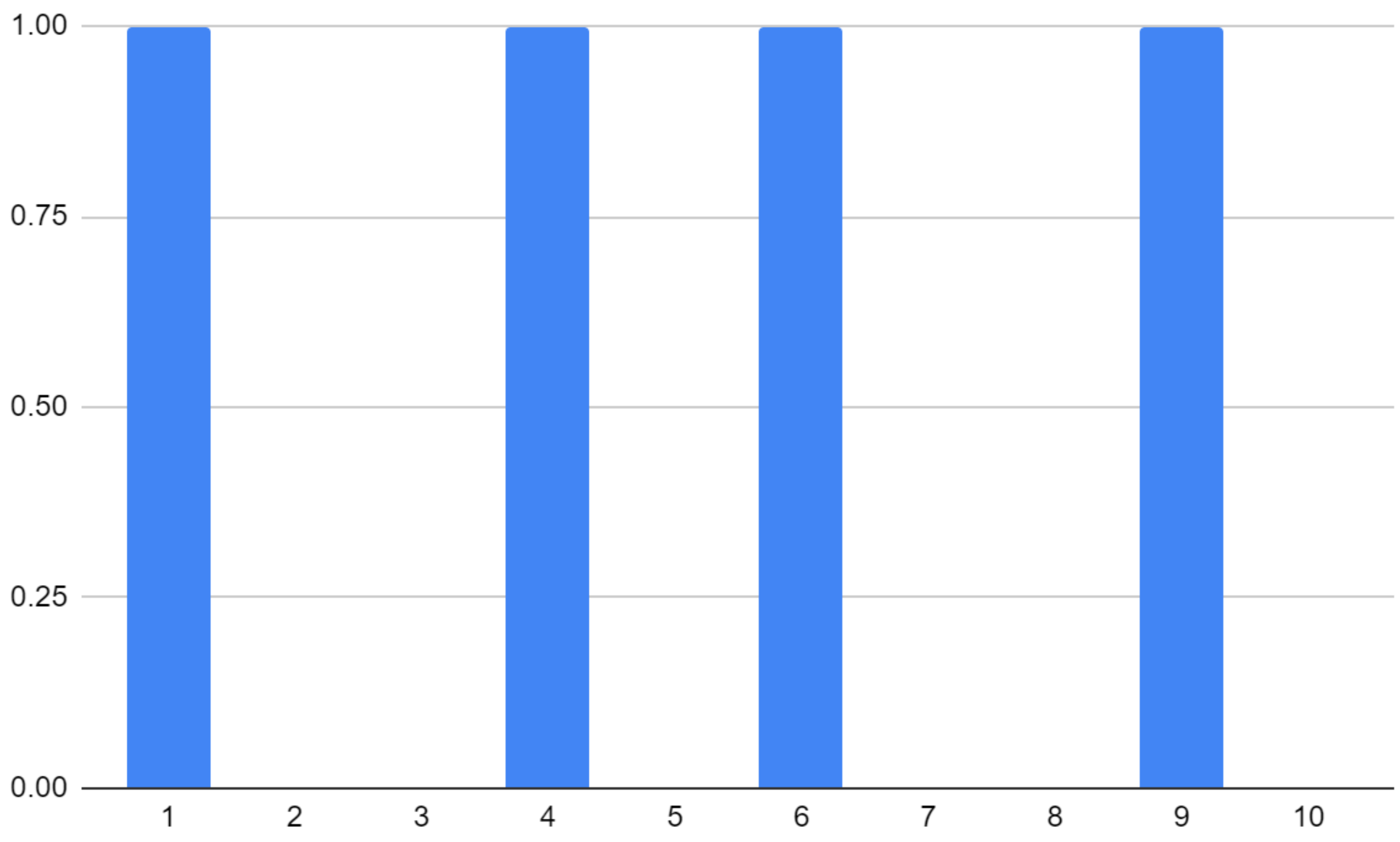
Parameter Estimation

- Data points representing the weight (in kgs) of students in a class.
- Whats mean and std deviation of the data ?
- Whats the probability that weight > 60



Parameter Estimation

- Data points representing the if it has rained or not in last 10 days.
- Whats the probability that it will rain tomorrow ?
- How many days will it rain in next 5 days ?

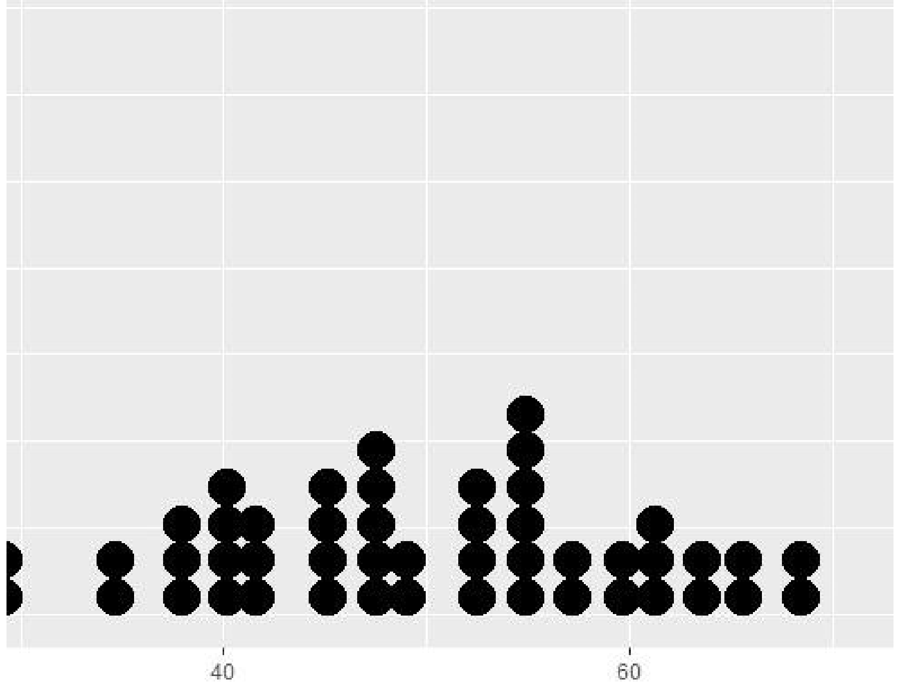


Parameter Estimation

- Any statistic used to estimate the value of an unknown parameter θ is called an estimator of θ.
    - mean and variance for Normal, rate (lambda) for Poisson, etc.

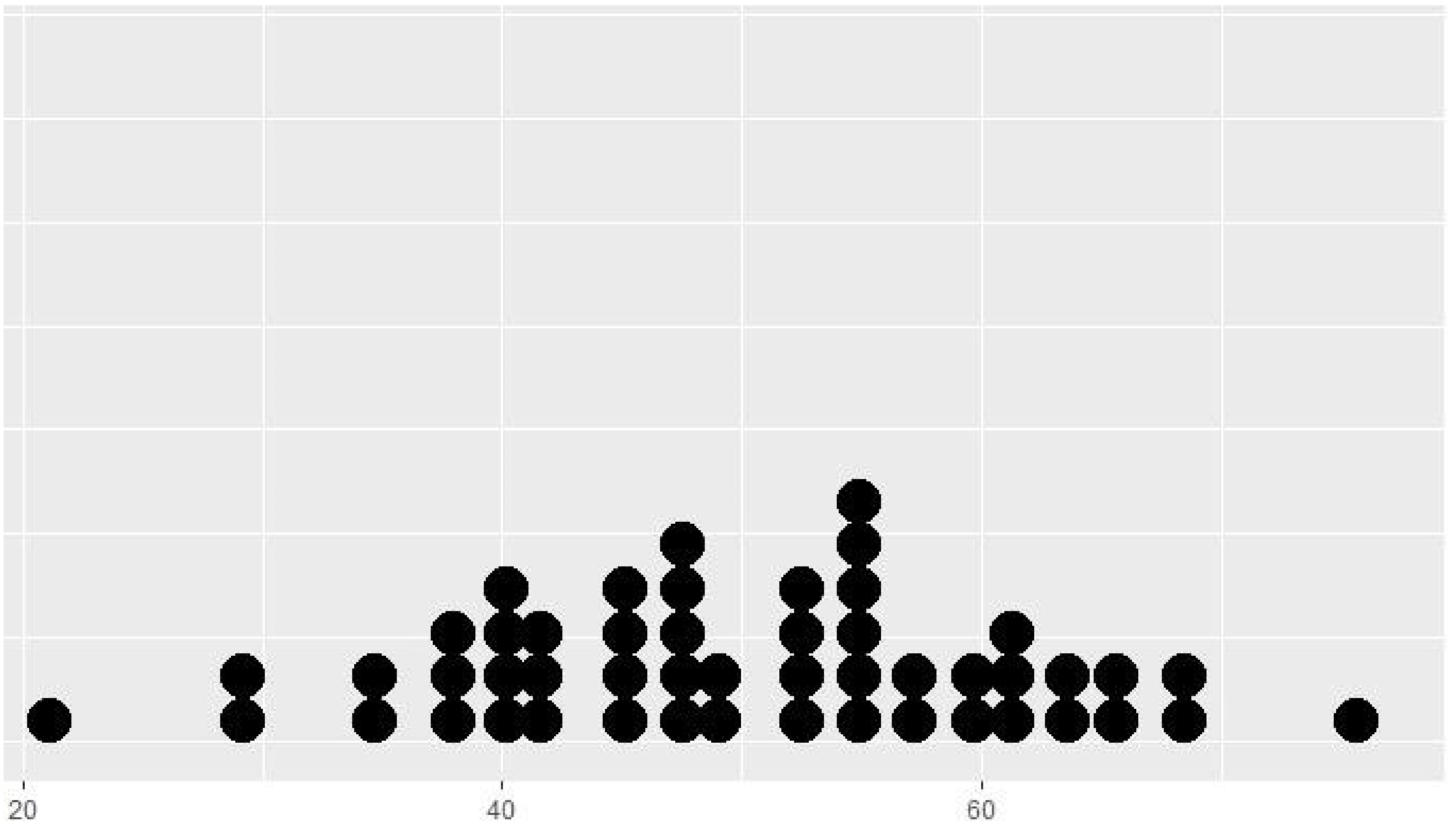
Parameter Estimation

- Any statistic used to estimate the value of an unknown parameter θ is called an estimator of θ.
  - mean and variance for Normal, rate (lambda) for Poisson, etc.
- **Maximum likelihood** estimator
- MLE can be defined as a method for estimating parameters of a distribution from sample data such that the likelihood of obtaining the observed data is maximized.
- Provides optimal way to fit a distribution to the data



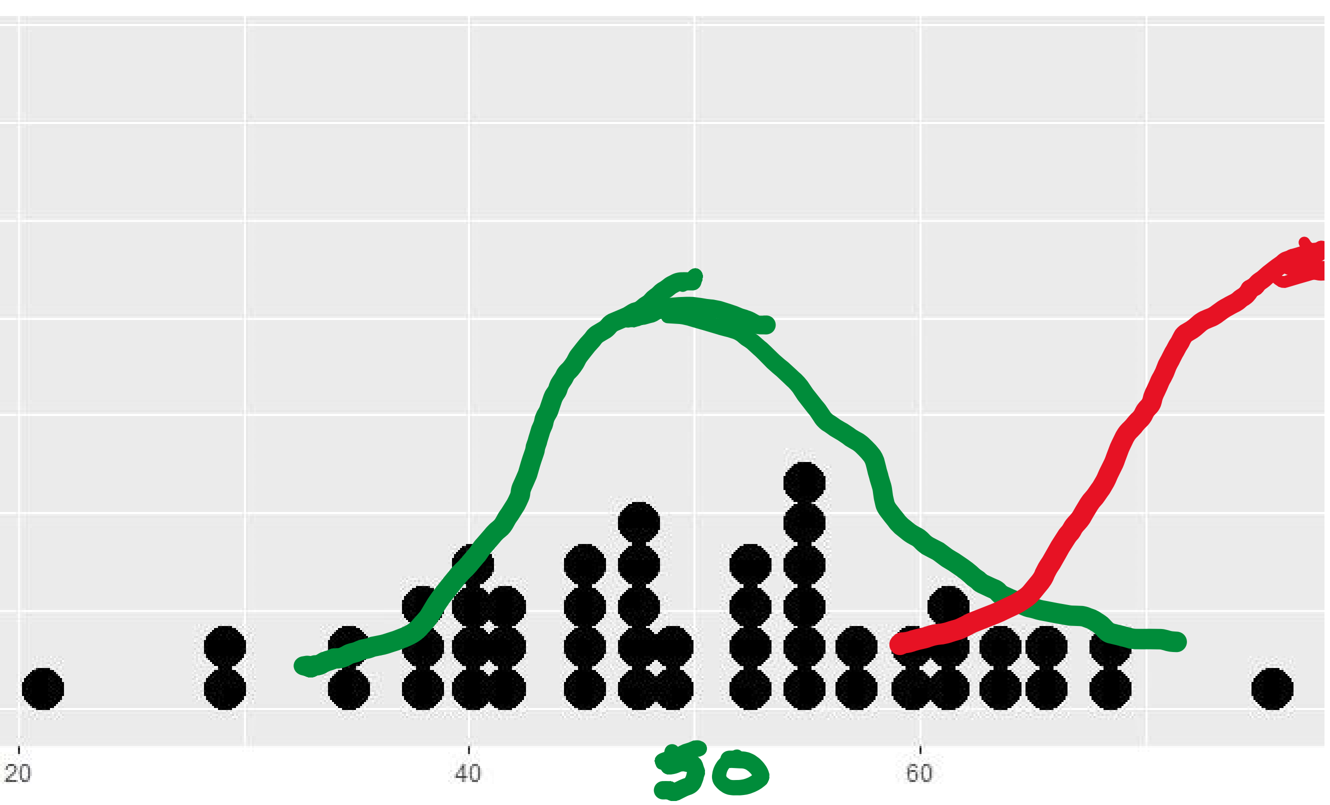
Parameter estimation

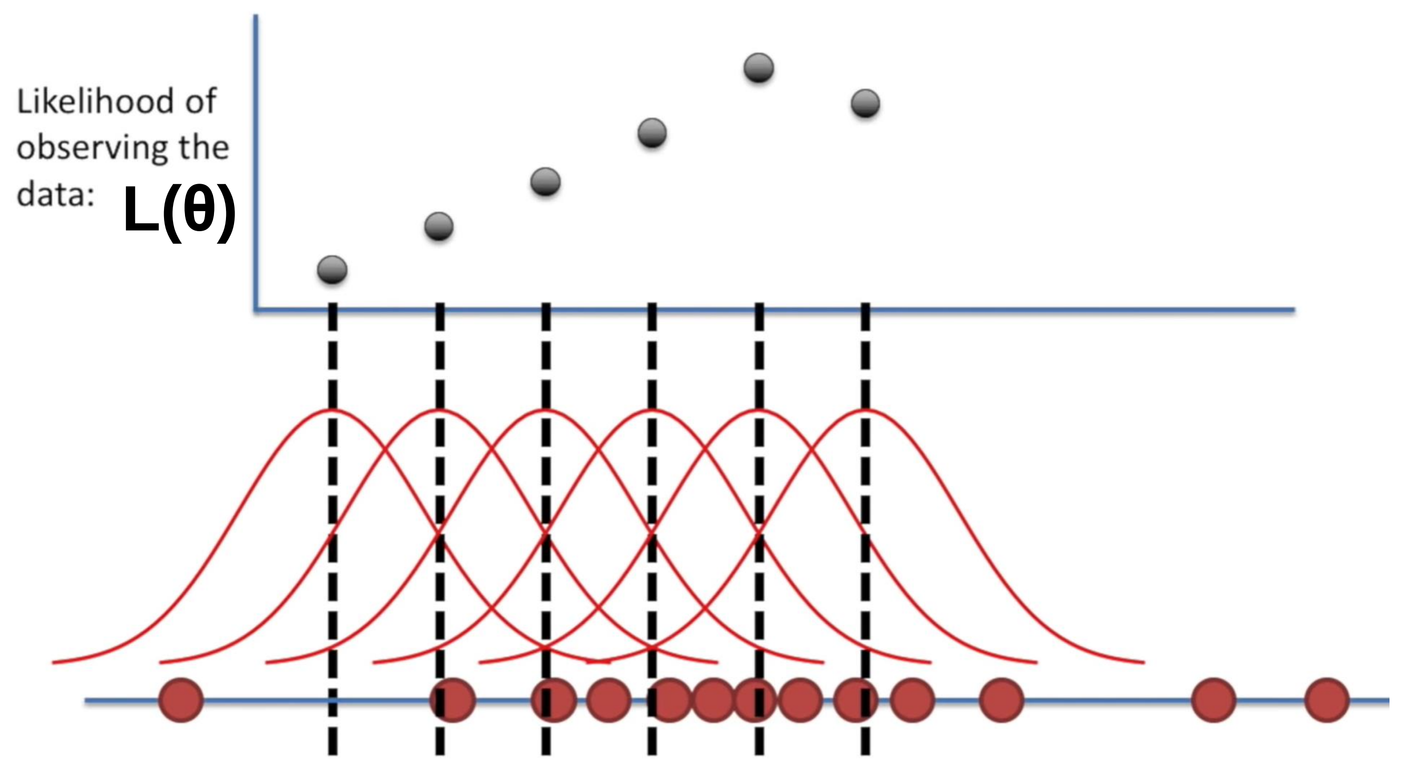
- which of the following would maximize the probability of observing the data
  - Mean = 100, SD = 10
  - Mean = 50, SD = 10



Parameter estimation

- which of the following would maximize the probability of observing the data
  - Mean = 100, SD = 10
  - Mean = 50, SD = 10





Parameter Estimation

Maximum likelihood estimator

- f(x1, . . . , xn|θ) represents the probability that the values x1, x2, . . . , xn will be observed when θ is the true value of the parameter

- Maximum Likelihood estimation : maximum likelihood estimate θ is defined to be that value of θ maximizing L(θ) = f (x1, . . . , xn|θ)

argmax_θ L(θ) = $f(x_1, \ldots, x_n|\theta)$ = argmax_θ $\log[f(x_1, \ldots, x_n|\theta)]$.

Note that L(θ) is not a distribution over θ but just a function of θ.

Independent and identically distributed (i.i.d.) assumption

$$f(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n)$$

Maximum likelihood estimation

- (Maximum Likelihood Estimator of a Bernoulli Parameter) Suppose you have data from n independent Bernoulli trials, $X1, \ldots, Xn$. Assuming the success probability is p what is the maximum likelihood estimator of p?

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases}$$

$$P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

$$P\{X_i = x\} = p^x(1-p)^{1-x}, \quad x = 0, 1$$

Maximum likelihood estimation

- (Maximum Likelihood Estimator of a Bernoulli Parameter) Suppose you have data from n independent Bernoulli trials, $X1, \ldots, Xn$. Assuming the success probability is p what is the maximum likelihood estimator of p?

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases}$$

$$P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

$$P\{X_i = x\} = p^x (1-p)^{1-x}, \quad x = 0, 1$$

$$\begin{aligned} f(x_1, \ldots, x_n | p) &= P\{X_1 = x_1, \ldots, X_n = x_n | p\} \\ &= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\ &= p^{\Sigma_1^n x_i}(1-p)^{n-\Sigma_1^n x_i}, \quad x_i = 0, 1, \quad i = 1, \ldots, n \end{aligned}$$

Maximum likelihood estimation

- (Maximum Likelihood Estimator of a Bernoulli Parameter) Suppose you have data from n independent Bernoulli trials, $X1, \ldots, Xn$. Assuming the success probability is p what is the maximum likelihood estimator of p?

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases}$$

$$P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

To determine the value of p that maximizes the likelihood,

$$\log f(x_1, \ldots, x_n | p) = \sum_1^n x_i \log p + \left(n - \sum_1^n x_i\right) \log(1 - p)$$

Maximum likelihood estimation

- (Maximum Likelihood Estimator of a Bernoulli Parameter) Suppose you have data from n independent Bernoulli trials, $X1, \dots, Xn$. Assuming the success probability is p what is the maximum likelihood estimator of p?

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases}$$

$$P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

To determine the value of p that maximizes the likelihood,

$$\log f(x_1, \dots, x_n | p) = \sum_1^n x_i \log p + \left(n - \sum_1^n x_i\right) \log(1 - p)$$

$$\frac{d}{dp} \log f(x_1, \dots, x_n | p) = \frac{\sum_1^n x_i}{p} - \frac{\left(n - \sum_1^n x_i\right)}{1 - p}$$

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

Maximum likelihood estimation

- (Maximum Likelihood Estimator of a Bernoulli Parameter) Suppose you have data from n independent Bernoulli trials, *X*1, . . . , *Xn*. Assuming the success probability is p what is the maximum likelihood estimator of p?

$$X_i = \begin{cases} 1 & \text{if trial } i \text{ is a success} \\ 0 & \text{otherwise} \end{cases} \qquad P\{X_i = 1\} = p = 1 - P\{X_i = 0\}$$

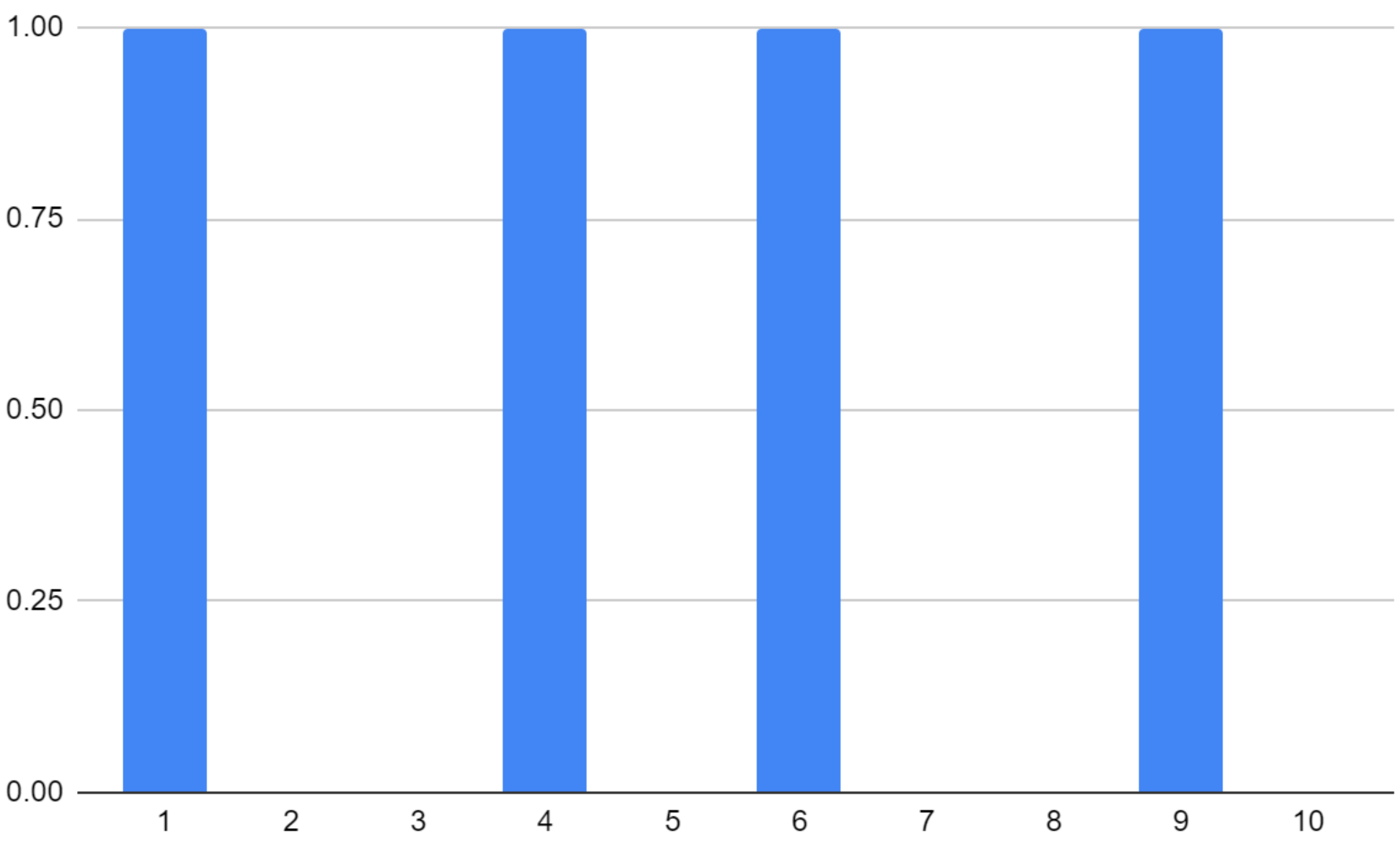
To determine the value of p that maximizes the likelihood,

proportion of the observed trials that result in successes.

$$\frac{d}{dp} \log f(x_1, \ldots, x_n | p) = \frac{\sum_1^n x_i}{p} - \frac{\left(n - \sum_1^n x_i\right)}{1 - p} \qquad \hat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

Suppose that each RAM (random access memory) chip produced by a certain manufacturer is, independently, of acceptable quality with probability *p*. Then if out of a sample of 1,000 tested 921 are acceptable, what is the maximum likelihood estimate of *p* ?

- Data points representing the if it has rained or not in last 10 days.
- Whats the probability that it will rain tomorrow ?
- How many days will it rain in next 5 days ?

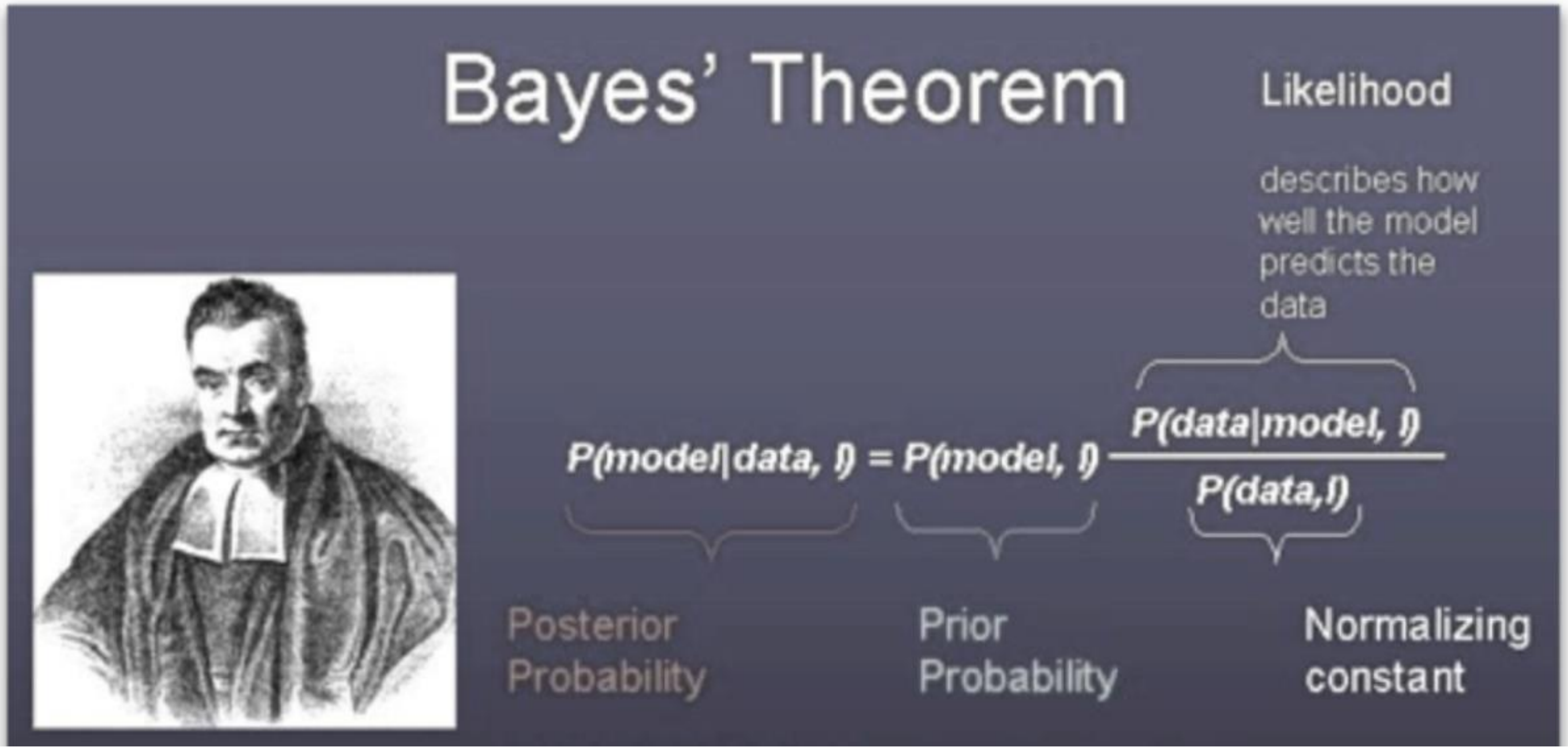


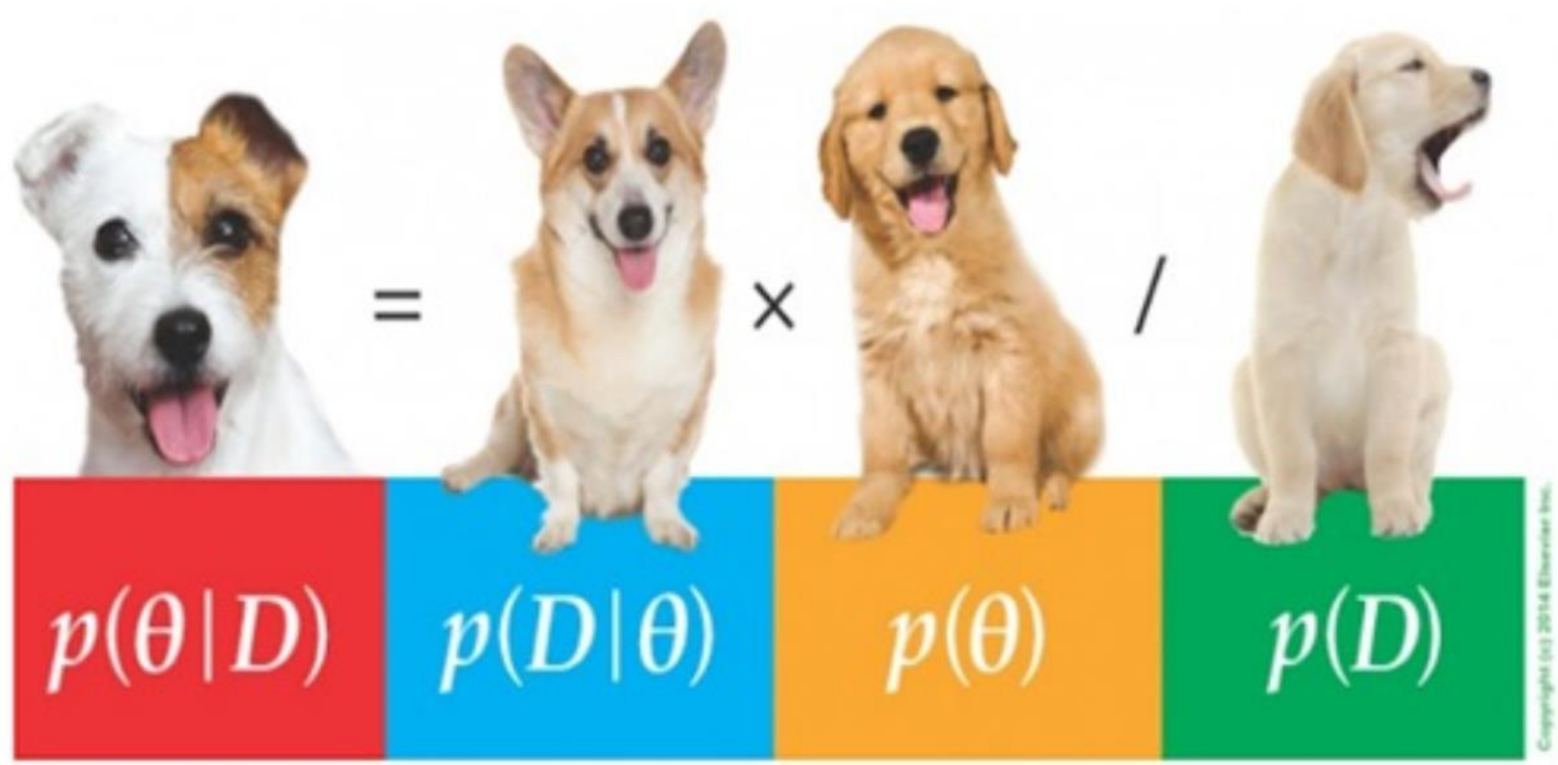
Maximum likelihood estimation

- (Maximum Likelihood Estimator of a Poisson Parameter) Suppose $X_1, \ldots, X_n$ are independent Poisson random variables each having mean λ. Determine the maximum likelihood estimator of λ.

Maximum likelihood estimation

- (Maximum Likelihood Estimator in a Normal Population) Suppose $X_1, \ldots, X_n$ are independent, normal random variables each with unknown mean μ and unknown standard deviation σ.


Bayes' Theorem
Likelihood
describes how well the model predicts the data
P(model|data, I) = P(model, I) P(data|model, I) / P(data,I)
Posterior Probability
Prior Probability
Normalizing constant


=
×
/
p(θ|D)
p(D|θ)
p(θ)
p(D)

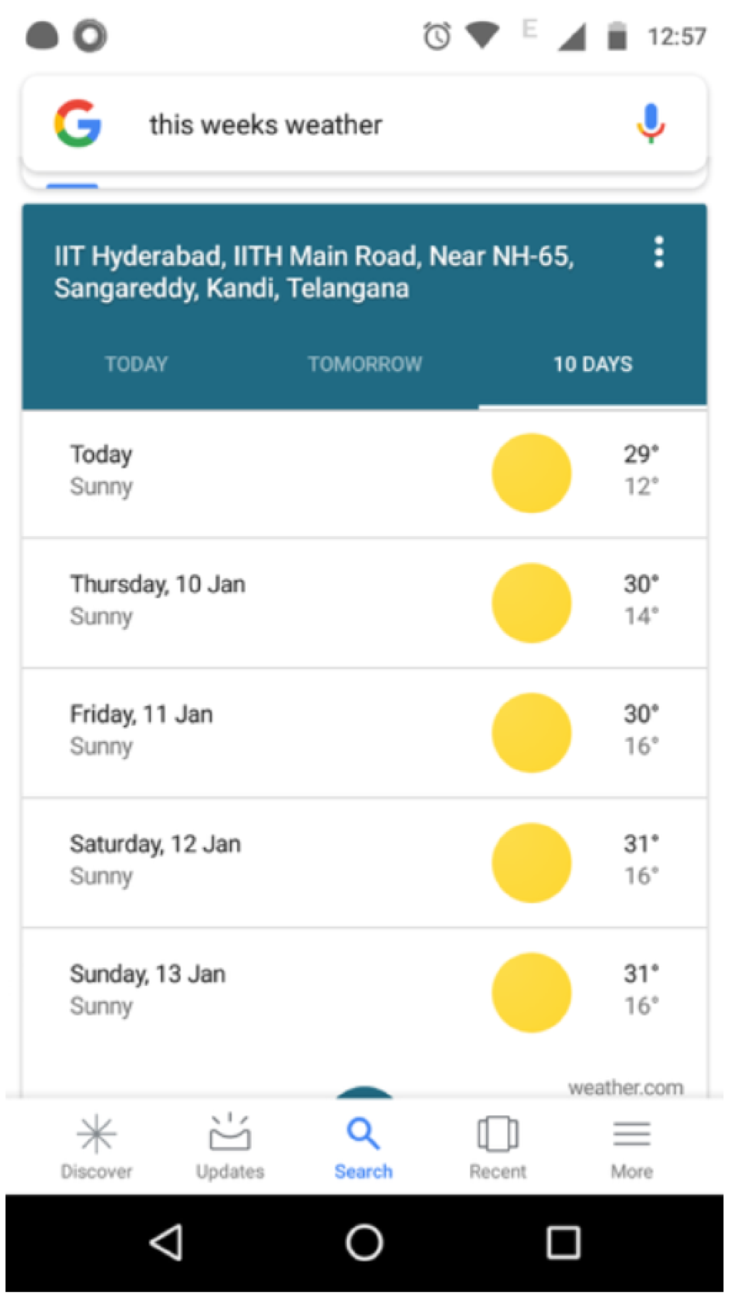
Rain prediction : ML estimation

X = [1,0,0, 0, 1, 0,….]

Model ?
Likelihood : p(X l model)
Learn model parameters : maximum likelihood (ML) estimation

$$\widehat{\Theta}_{ML} = \underset{\Theta}{\operatorname{argmax}} \; \mathcal{L}$$

$$\mathcal{L} = \sum_{\mathrm{x}_i \in \mathcal{X}} log\ prob(\mathrm{x}_i | \Theta)$$



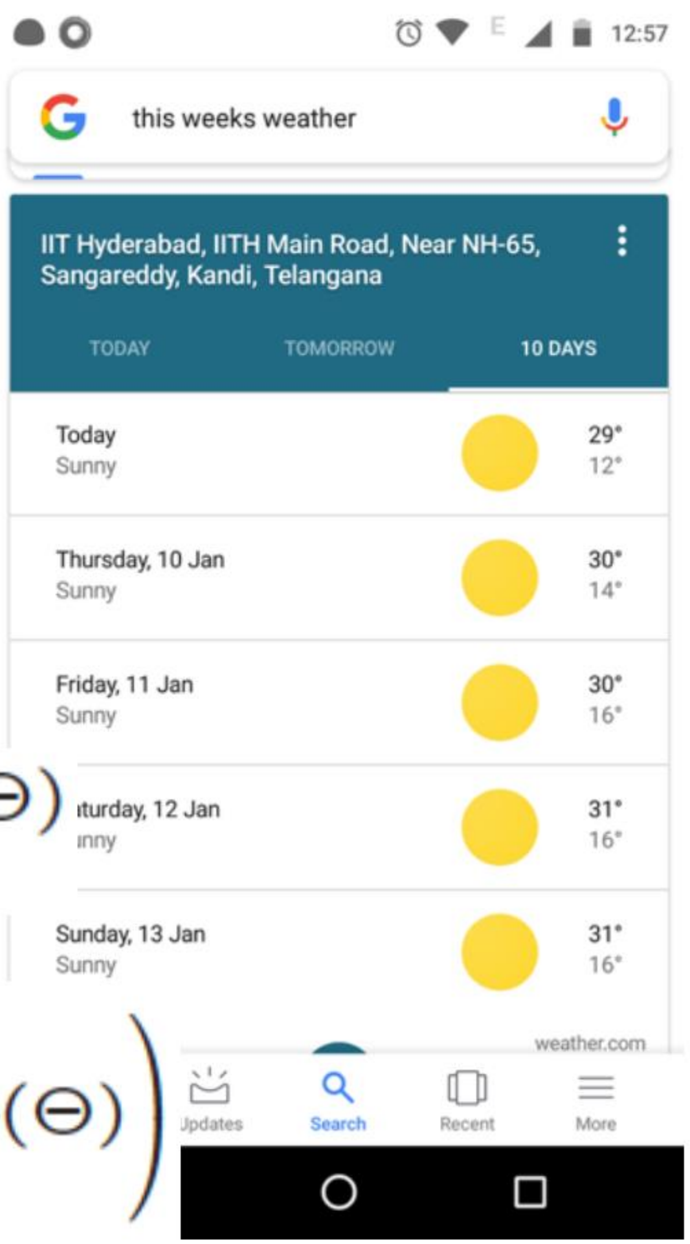
Rain prediction : MAP estimation

Maximum A-Posteriori (MAP) estimation

$$prob(\Theta|\mathcal{X}) = \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{prob(\mathcal{X})}$$

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \prod_{\mathrm{x}_i \in \mathcal{X}} prob(\mathrm{x}_i|\Theta) \cdot prob(\Theta)$$

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left(\sum_{\mathrm{x}_i \in \mathcal{X}} log\ prob(\mathrm{x}_i|\Theta) + log\ prob(\Theta) \right)$$



Seek that value for θ which maximizes the posterior prob(θ|X).

What Does the MAP Estimate Get Us That the ML Estimate Does NOT

- MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in θ

- MAP estimation “pulls” the estimate toward the prior. The more focused our prior belief, the larger the pull toward the prior.

- “smoothing” role (Laplace smoothing) for parameter estimation.

MAP Estimation Example [Election]

- Consider a survey where voters were asked if they will vote for NDA or UPA in the next election. Let p be the probability that an individual will vote NDA.
- xi is either NDA or UPA, and p is the probability some body votes for NDA
- nd is the number of individuals who are planning to vote NDA

$$\widehat{p}_{ML} = \frac{n_d}{N}$$

N= 20 and if 12 out of 20 said that they were going to vote NDA, we get the following the ML estimate for p: ˆpML = 0.6.

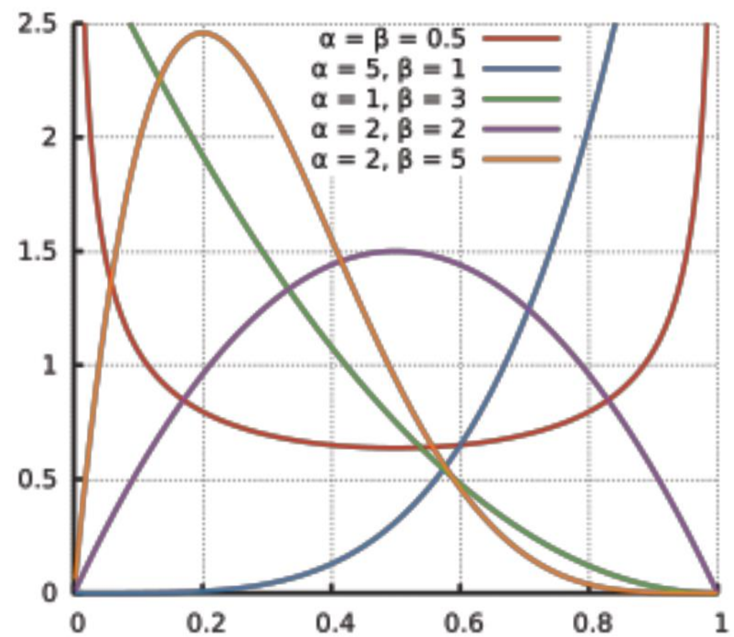
MAP Estimate : Prior Belief on p

- The prior for p must be zero outside the [0, 1] interval.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the [0, 1] interval.
- beta distribution that is parameterized by two “shape” constants α and β does the job nicely for expressing our prior beliefs concerning p:

$$prob(p|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

$$B = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

mode $\dfrac{\alpha-1}{\alpha+\beta-2}$



Example [Indian Election]

- Due to some emerging situations, lets assume that voters are equally likely to vote for UPA and NDA.
- Consider a prior distribution for p that has a peak at 0.5. Setting α = β=5 gives us a distribution for p that has a peak in the middle of the [0, 1] interval.

$$\hat{p}_{MAP} = \operatorname*{argmax}_{p} \left(\sum_{x \in \mathcal{X}} log\ prob(x|p)\ +\ log\ prob(p) \right)$$

MAP Estimation [Indian Election]

$$\hat{p}_{MAP} = \underset{p}{\text{argmax}} \Big(\; n_d \cdot log\ p + (N - n_d) \cdot log\ (1-p) \; + log\ prob(p) \; \Big)$$

$$\frac{n_d}{p} - \frac{(N - n_d)}{(1-p)} + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1-p} = 0$$

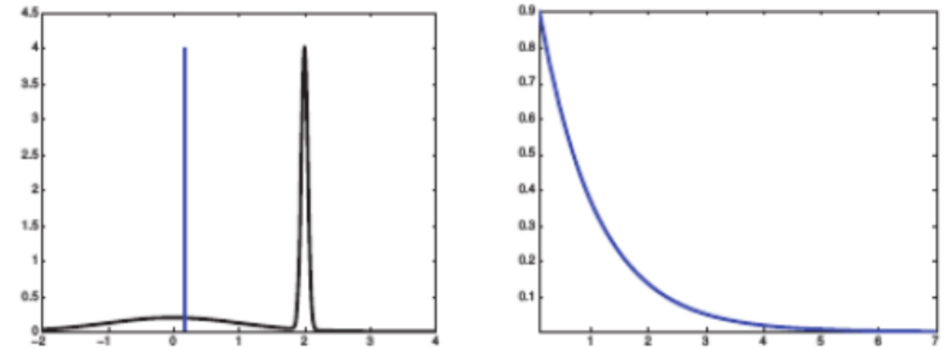
$$\hat{p}_{MAP} = \frac{n_d + \alpha - 1}{N + \alpha + \beta - 2}$$

$$= \frac{n_d + 4}{N + 8}$$

With N = 20 and with 12 of the 20 previously saying they would vote NDA, the MAP estimate forp is 0.571 with α and β both set to 5.

Rain prediction : Bayesian estimation

Both ML and MAP return only single and specific values for the parameter Θ.

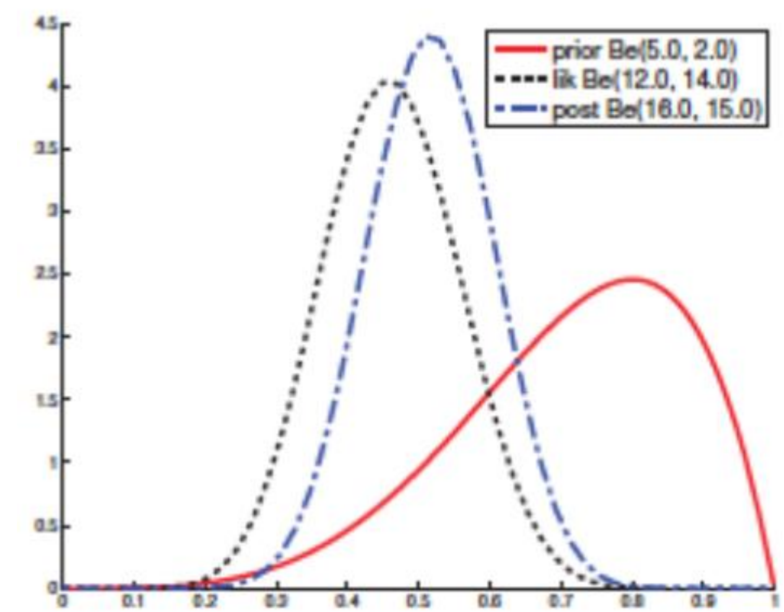


(a) (b)

$$prob(\Theta|\mathcal{X}) \quad = \quad \frac{prob(\mathcal{X}|\Theta) \ \cdot \ prob(\Theta)}{prob(\mathcal{X})}$$

$$prob(\mathcal{X}) \ = \ \int_{\Theta} prob(\mathcal{X}|\Theta) \cdot prob(\Theta) \ d\Theta$$

$$prob(\tilde{\mathrm{x}}|\mathcal{X}) \quad = \quad \int_{\Theta} prob(\tilde{\mathrm{x}}|\Theta) \cdot prob(\Theta|\mathcal{X}) \ d\Theta$$



Posterior is a “compromise” between the prior and likelihood.
posterior mean is convex combination of the prior mean and the MLE

: $\lambda m_1 + (1-\lambda)\hat{\theta}_{MLE}$


THANK YOU