A Primer to Probability distribution and Parameter Estimation

Mean and Variance of Random Variable

- Expected value of a random variable X is the long-run average value of repetitions of the experiment

$$\mathrm{E}[X] = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k \,.$$

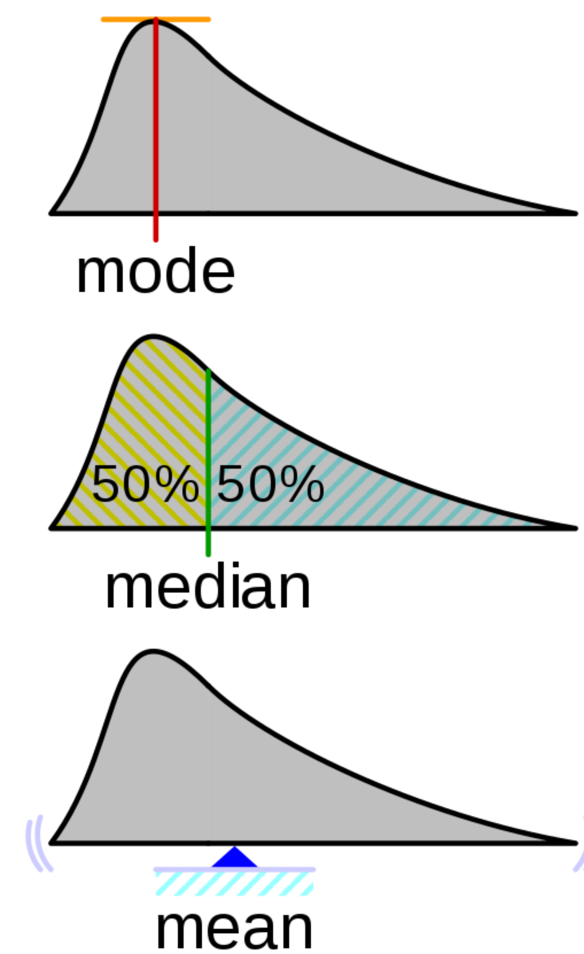
- Variance : Spread of the random variable values

$$\mathrm{Var}(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$$

$\sqrt{\mathrm{Var}(X)}$ is called the *standard deviation* of X.

$$W = 0 \quad \text{with probability } 1$$

$$Y = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases} \qquad Z = \begin{cases} -100 & \text{with probability } \frac{1}{2} \\ 100 & \text{with probability } \frac{1}{2} \end{cases}$$



Common Discrete Distributions
Bernoulli and Binomial

- Let $X \in \{0, 1\}$ be a binary random variable, with probability of "success" θ, X has a Bernoulli distribution, $X \sim \text{Ber}(\theta)$

  E.g Coin toss, Rain or not

$$\text{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$$

$$\text{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

Common Discrete Distributions
Bernoulli and Binomial

- Let $X \in \{0, 1\}$ be a binary random variable, with probability of "success" θ, X has a Bernoulli distribution, $X \sim \mathrm{Ber}(\theta)$

  E.g Coin toss, Rain or not

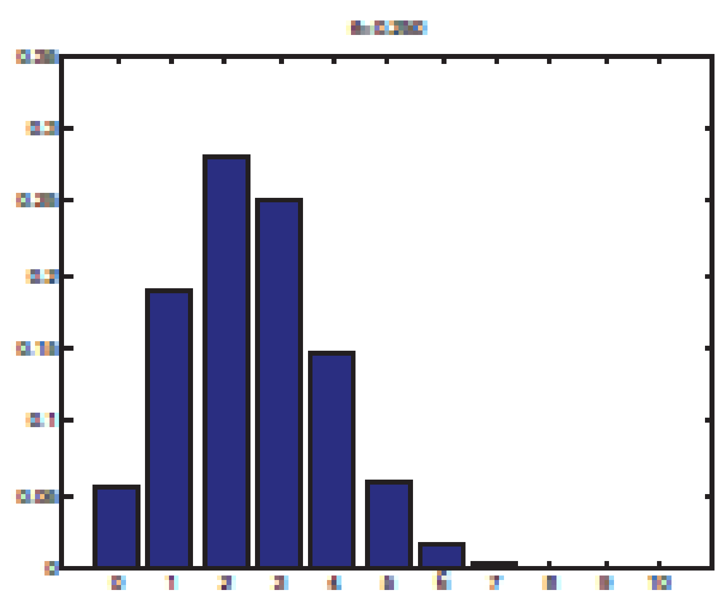
$$\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)} \qquad \mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

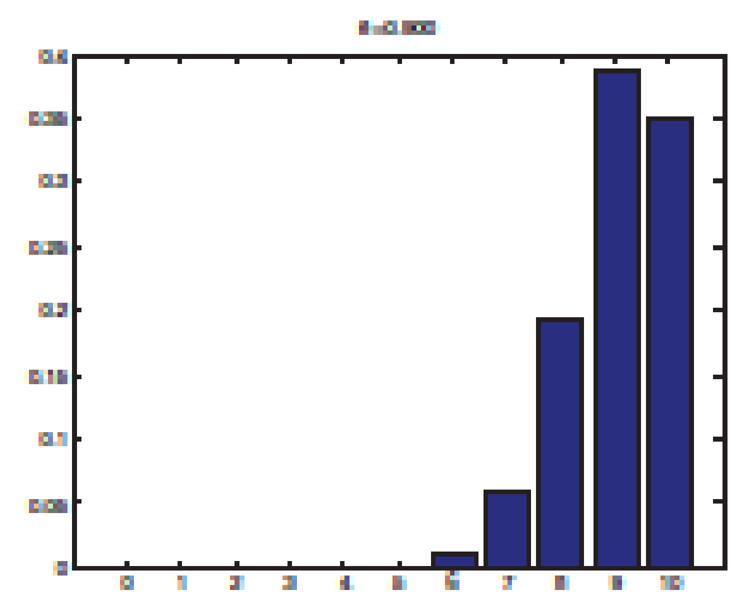
- Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a binomial distribution, written as $X \sim \mathrm{Bin}(n, \theta)$.

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

$$\text{mean} = n\,\theta, \quad \text{var} = n\theta(1-\theta)$$

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$





Discrete Distributions : Multinoulli, Multinomial

- Model the outcomes of tossing a K -sided die : categorical/Multinoulli distribution,

  $x \sim \mathrm{Cat}(\boldsymbol{\theta})$, $p(x = j|\boldsymbol{\theta}) = \theta_j$.

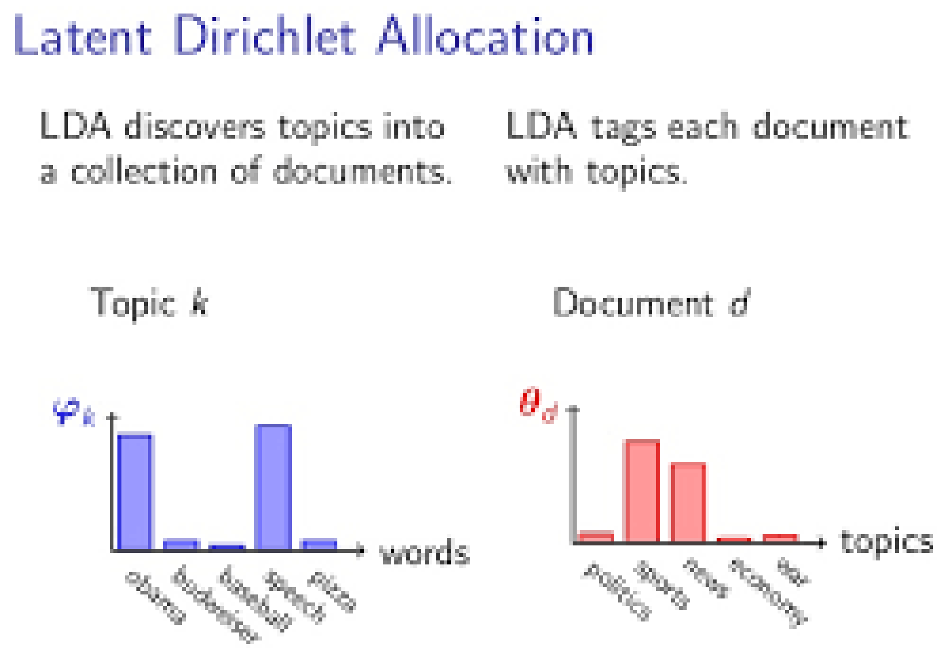
- Multinomial distribution : Models the outcome of n dice rolls, let x = (x1, . . . , xk) be a random vector, where xj number of times side j of the die occurs.

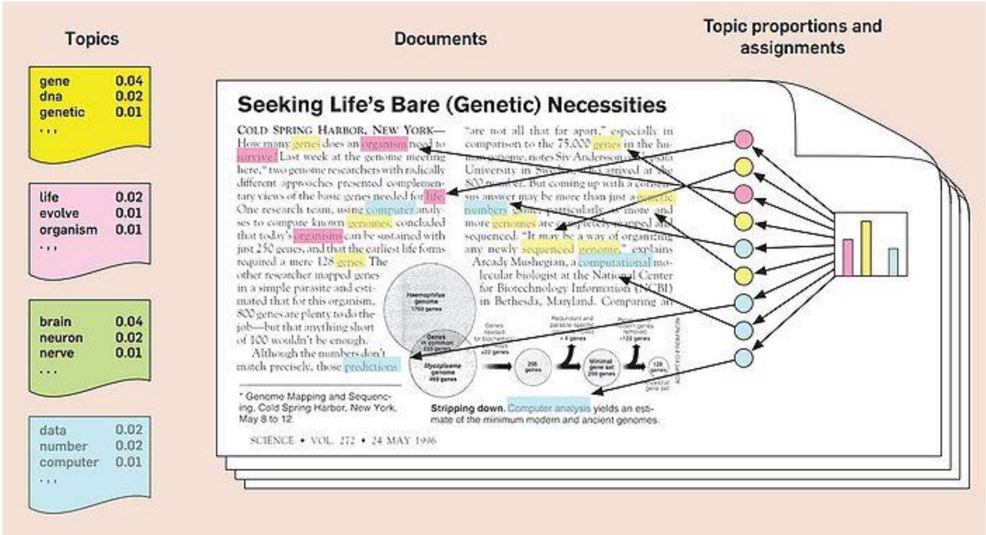
$$\mathrm{Mu}(\mathbf{x}|n,\boldsymbol{\theta}) \triangleq \binom{n}{x_1 \ldots x_K} \prod_{j=1}^{K} \theta_j^{x_j}$$

$$\mathrm{Cat}(x|\boldsymbol{\theta}) \triangleq \mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta}) \qquad \mathrm{Mu}(\mathbf{x}|1,\boldsymbol{\theta}) = \prod_{j=1}^{K} \theta_j^{\mathbb{I}(x_j=1)}$$

- Probabilistic topic model

- Text classification





Poisson distribution

"My husband always loves your Poisson distribution – it's something to do with him being a mathematician."

- Model number of events occurring in a fixed interval of time/space

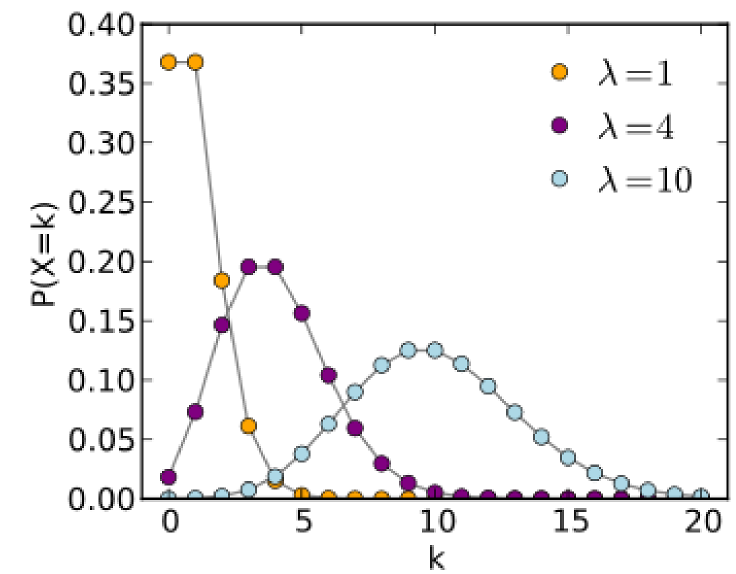
$$P(k \text{ events in interval}) = e^{-\lambda} \frac{\lambda^k}{k!}$$

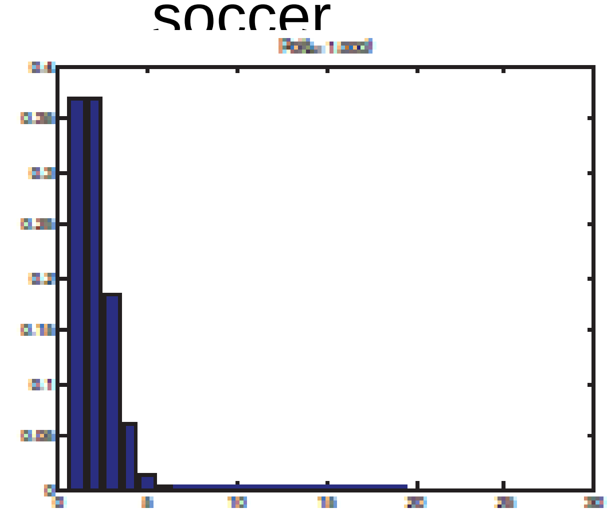
- λ is the average (mean) number of events per interval, k = 0, 1, 2, …, events occur independently, rate is a constant.
- Models rare events
  - Number of misprints on a page of a book.
  - average number of goals in a World Cup soccer match is approximately 2.5 ; λ = 2.5.

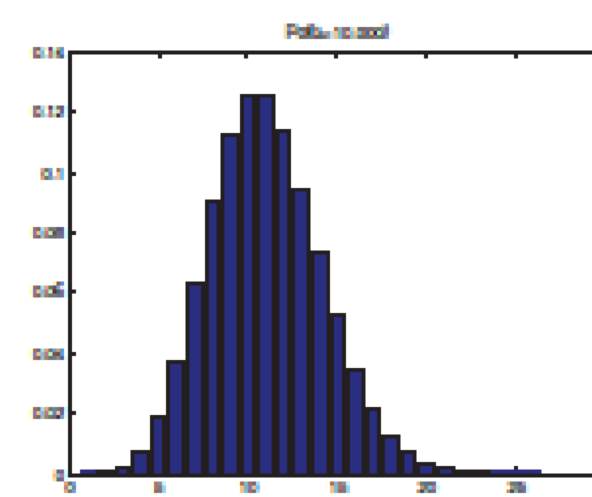
$$P(k \text{ goals in a match}) = \frac{2.5^k e^{-2.5}}{k!}$$

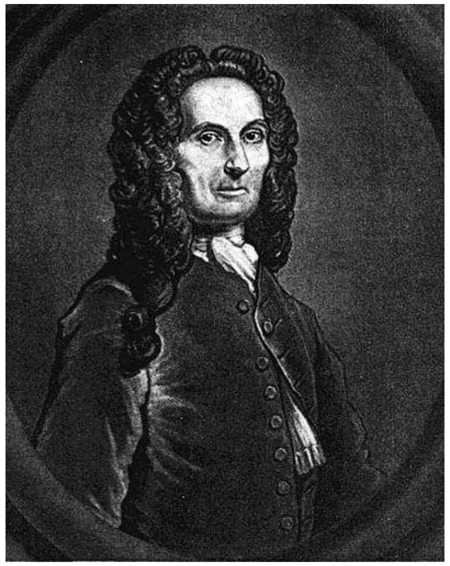
  - Number of wrong telephone numbers

that are dialed in a day.

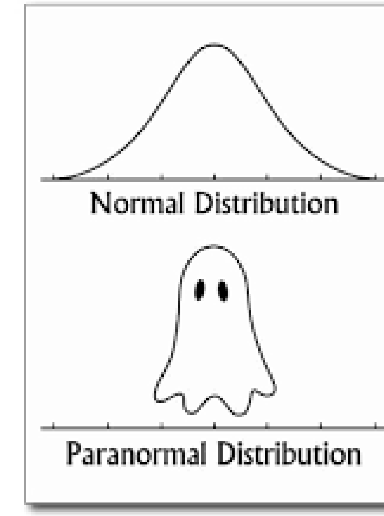








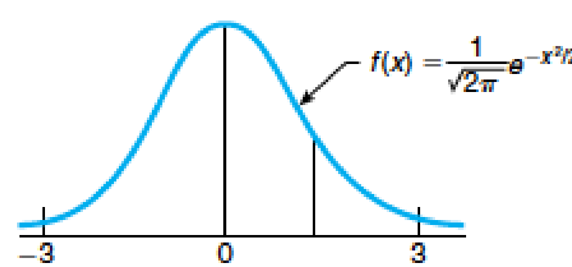
Normal/Gaussian Random Variables

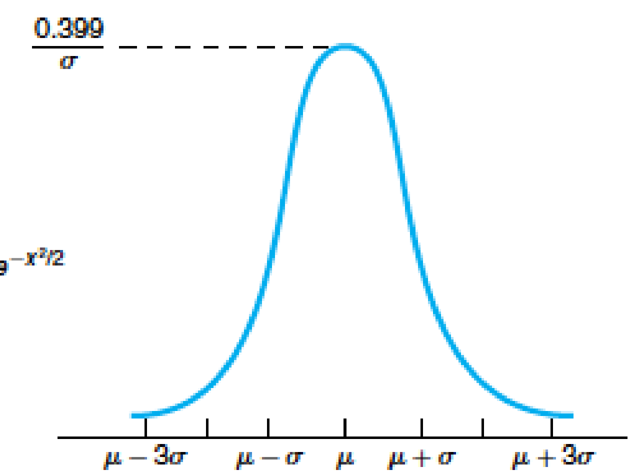


- 1809 Gauss published his monograph "Theoria motus corporum coelestium in sectionibus conicis solem ambientium"
- *All distributions of frequency other than normal are ‘abnormal’*- Pearson
- A random variable is said to be normally distributed

  with parameters μ and σ2, X ~ N(μ, σ2)

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty^*$$

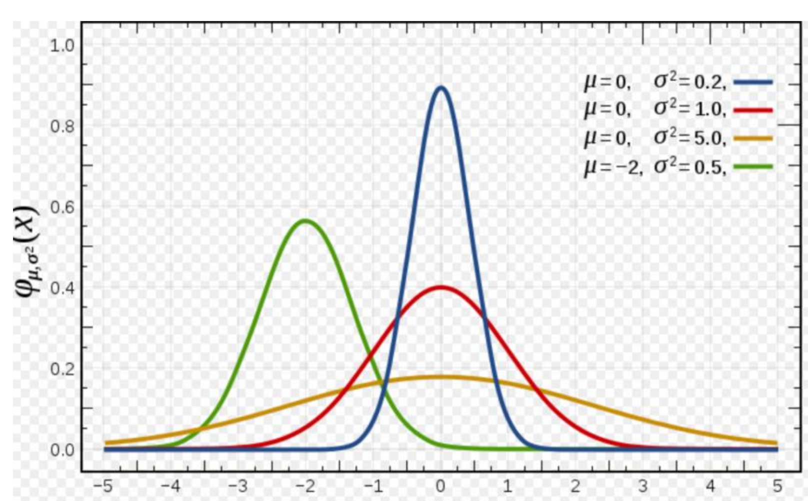


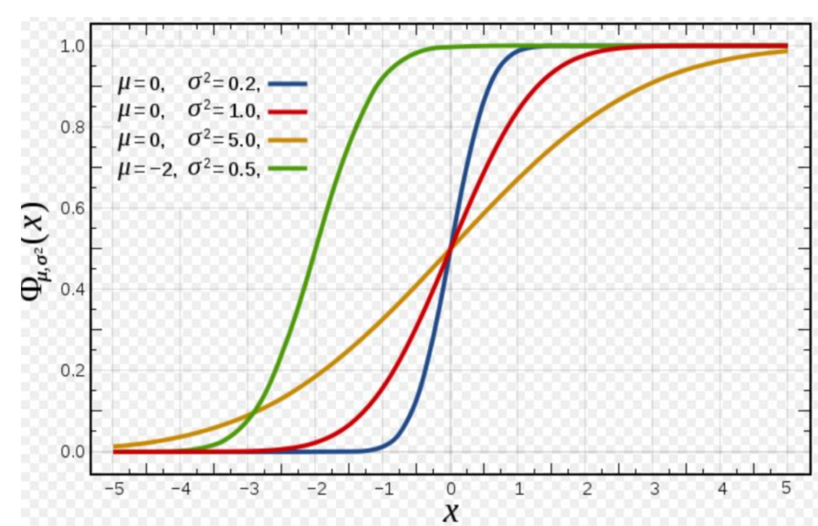


- μ = E [X] is the mean (and mode), and σ2 = var [X] is the variance.

$$\Phi(x; \mu, \sigma^2) \triangleq \int_{-\infty}^{x} \mathcal{N}(z|\mu, \sigma^2) dz$$

- CDF of the Gaussian



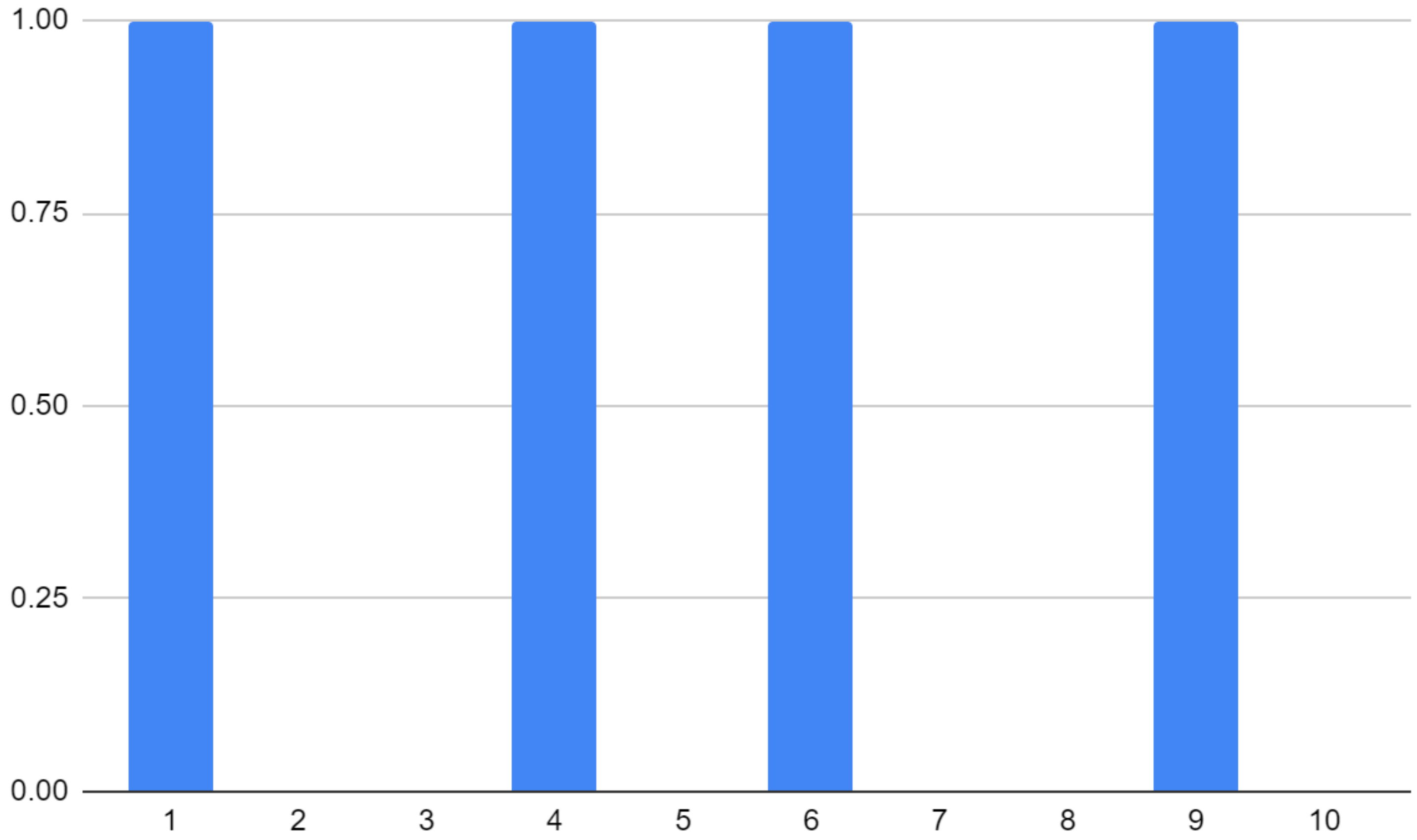


Probability Distribution Summary

- X : Discrete
  - Binary valued scalar (0/1) : Bernoulli
  - Binary valued vector (one of K): Multinoulli/categorical
  - Multivalued scalar (M of N): Binomial
  - Multivalued vector (M1, M2, … MK) : Multinomial
  - Integer valued scalar (1 to infinity) : Poisson
- X : continous, real valued
  - Interval [a,b] : Uniform, Interval [0,1] : Beta
  - non-negative (0,infinity) : Exponential, Gamma
  - real line (-infinity, infinity) : Normal, students, Laplace
  - Vector : Real valued : Gaussian ; Simplex : Dirichlet

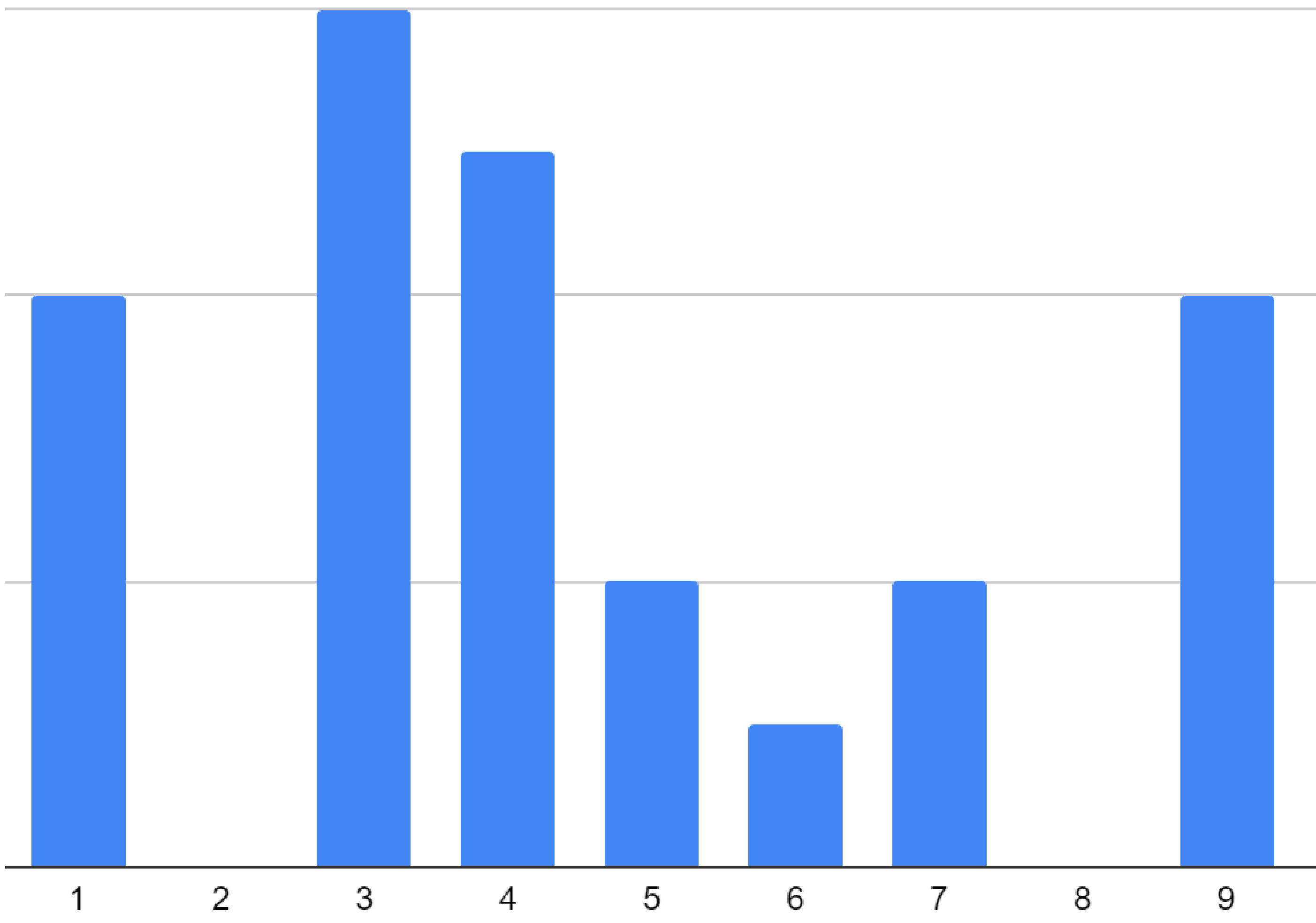
Parameter Estimation

- 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 ?



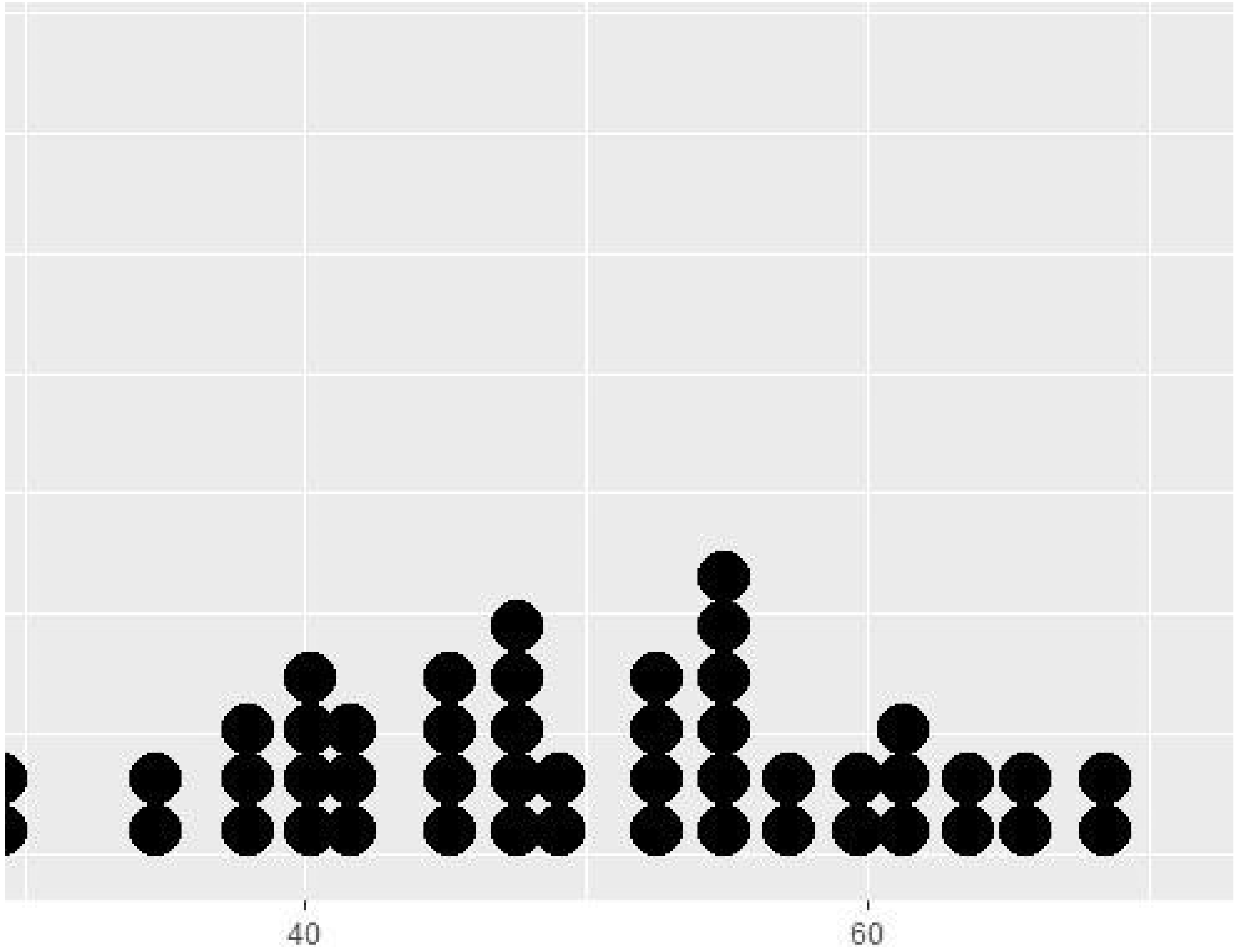
Parameter Estimation

- The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:
- 4, 0, 6, 5, 2, 1, 2, 0, 4, 3
- Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.



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

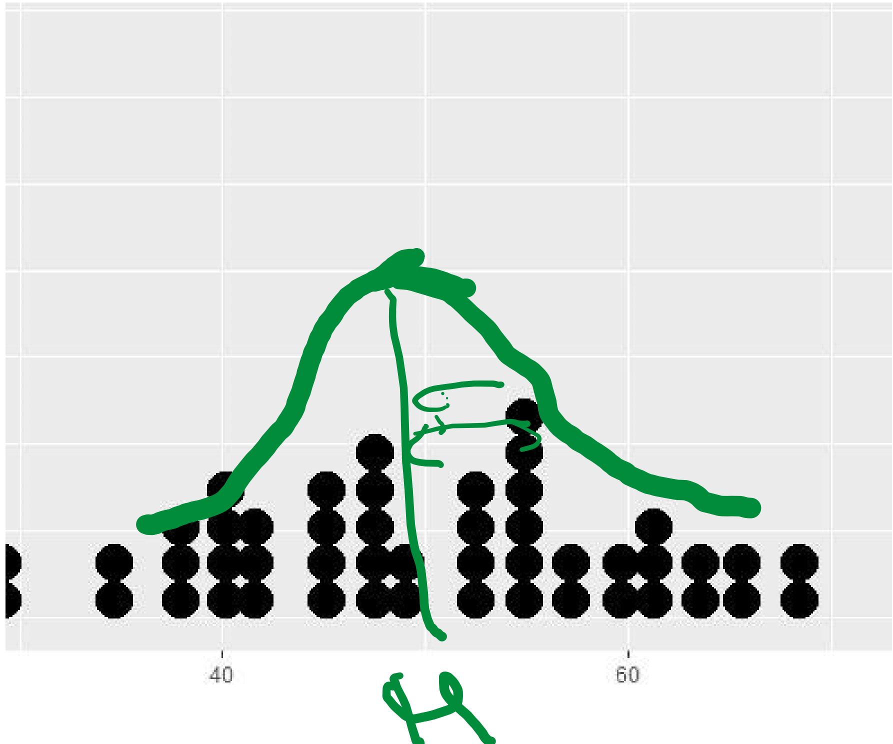


Parameter Estimation

- 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

$X \sim N(\mu, \sigma^2)$

$\theta = (\mu, \sigma)$

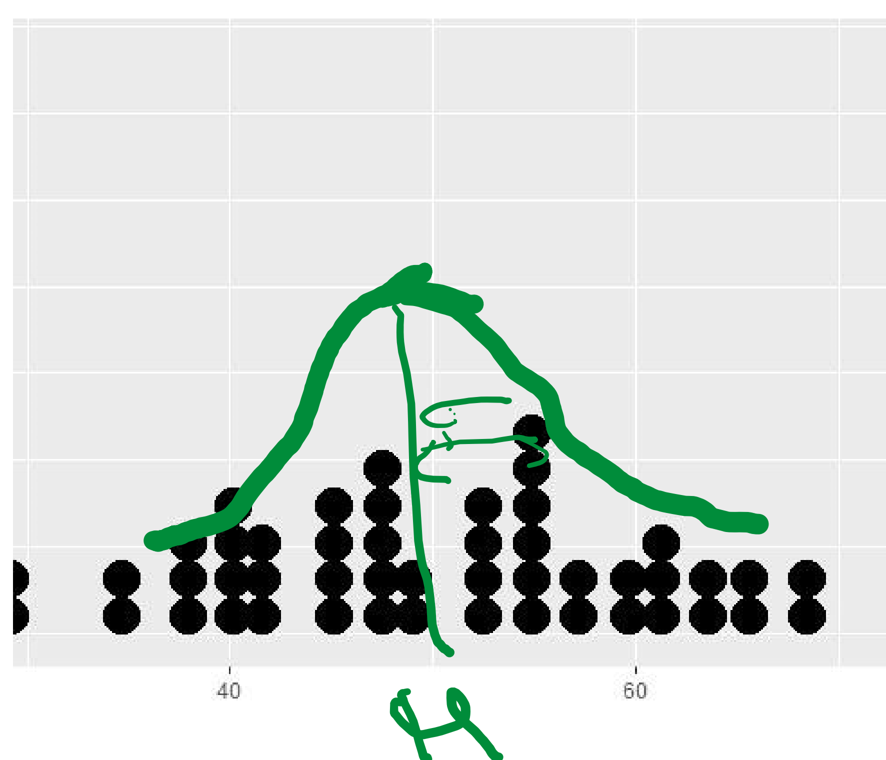


Parameter Estimation

- 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

$X \sim N(\mu, \sigma^2)$

$\theta = (\mu, \sigma)$



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 ˆθ 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 $\log[f(x_1, \ldots, x_n|\theta)]$.

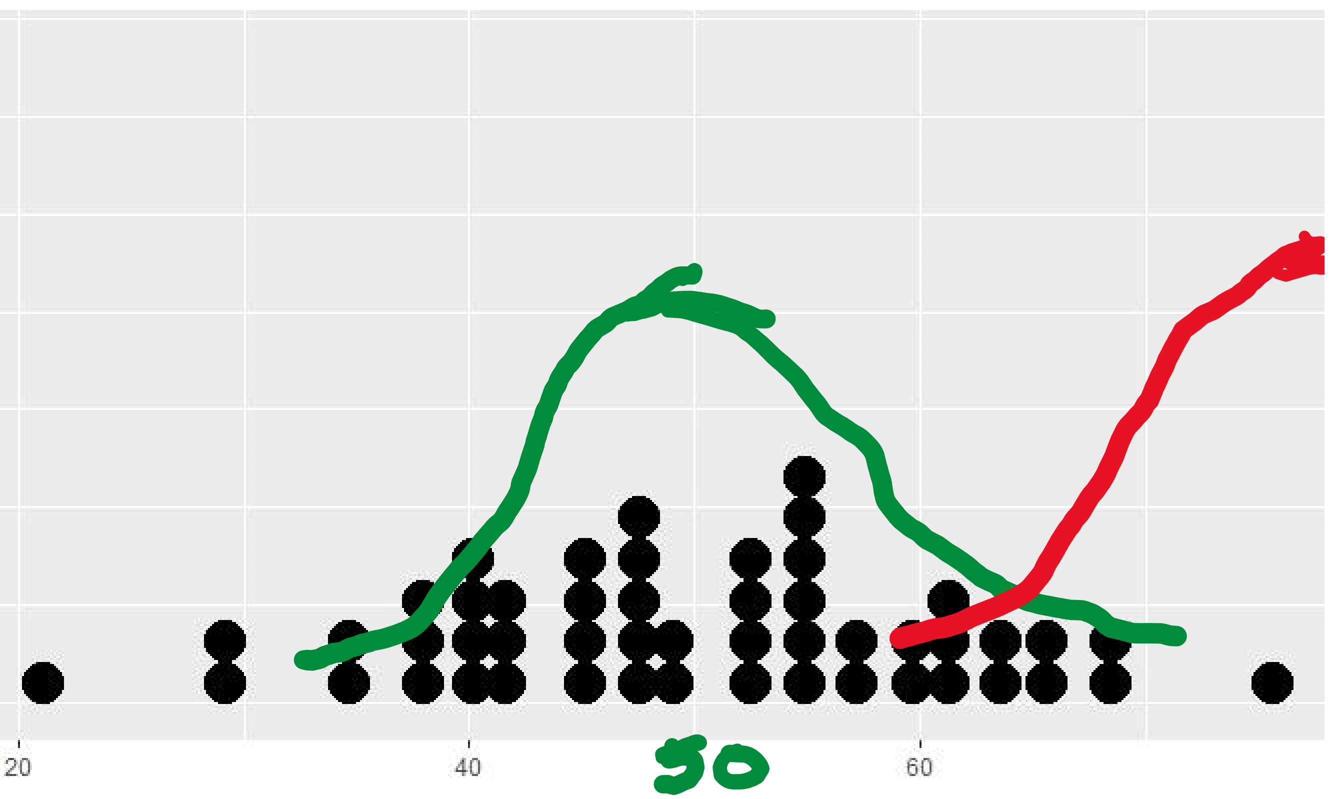
Note that L(θ) is not a distribution over θ but just a function of θ.

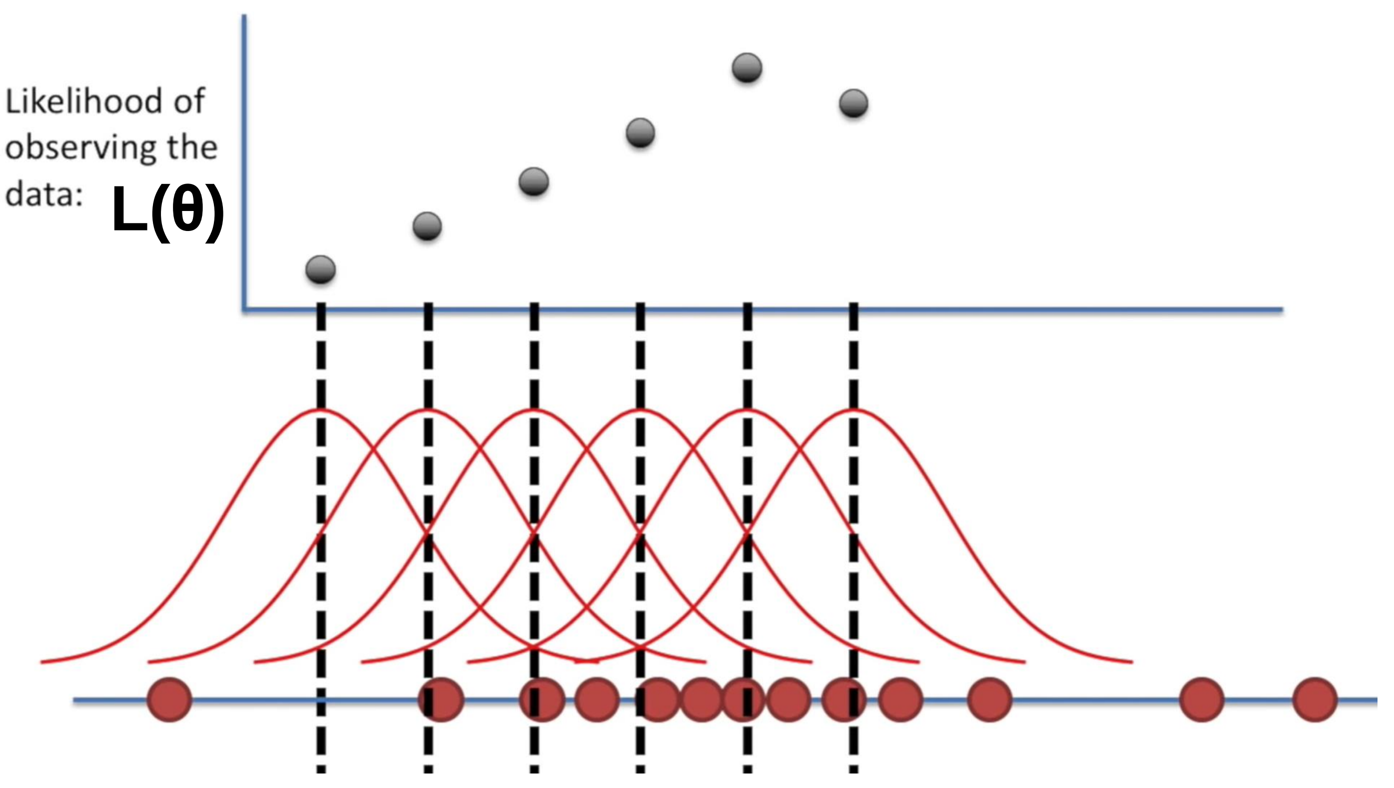
Independent and identically distributed (i.i.d.) assumption

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

Parameter estimation

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





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, *X*1, . . . , *Xn*. Assuming the success probability is p what is the maximum likelihood estimator of p?

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

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

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, $X1, \ldots, Xn$. Assuming the success probability is p what is the maximum likelihood estimator of p?

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

To determine the value of p that maximizes the likelihood,

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

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

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.

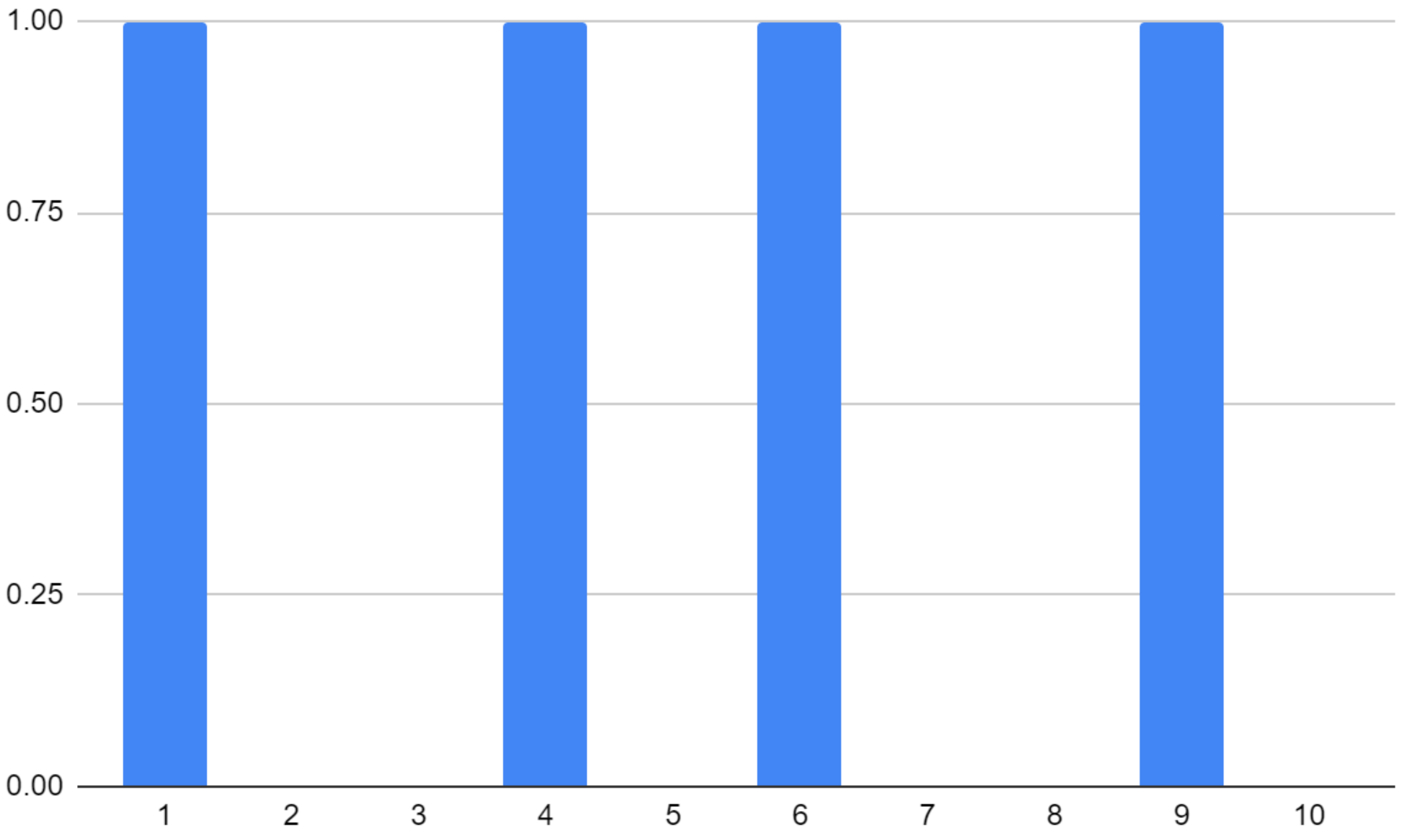
$$\frac{d}{dp} \log f(x_1, \ldots, x_n | p) = \frac{\sum_1^n x_i}{p} - \frac{\left(n - \sum_1^n x_i\right)}{1-p} \qquad \hat{p} = \frac{\sum_{i=1}^n x_i}{n}$$

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 ?



Parameter Estimation

- Multinomial
- 3,1,2,4,3,5,6,1,3,4

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \mathbf{x} = (0,0,1,0,0,0)^{\mathrm{T}}.$$

data set $\mathcal{D}$ of N independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}. \qquad m_k = \sum_n x_{nk}$$

Parameter Estimation

- Multinomial

data set $\mathcal{D}$ of N independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}.$$

Parameter Estimation : Multinomial

data set $\mathcal{D}$ of N independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_N$.

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}.$$

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

Constraint

$$\mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

ML solution !

MLE : Poisson !

- (Maximum Likelihood Estimator of a Poisson Parameter) Suppose X1, . . . , Xn are independent Poisson random variables each having mean λ. Determine the maximum likelihood estimator of λ.

$$f(x_1, \ldots, x_n|\lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$= \frac{e^{-n\lambda}\lambda^{\Sigma_1^n x_i}}{x_1! \ldots x_n!}$$

$$\log f(x_1, \ldots, x_n|\lambda) = -n\lambda + \sum_1^n x_i \log\lambda - \log c$$

MLE : Poisson !

- (Maximum Likelihood Estimator of a Poisson Parameter) Suppose X1, . . . , Xn are independent Poisson random variables each having mean λ. Determine the maximum likelihood estimator of λ.

$$\frac{d}{d\lambda} \log f(x_1, \ldots, x_n|\lambda) = -n + \frac{\sum_{1}^{n} x_i}{\lambda}$$

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

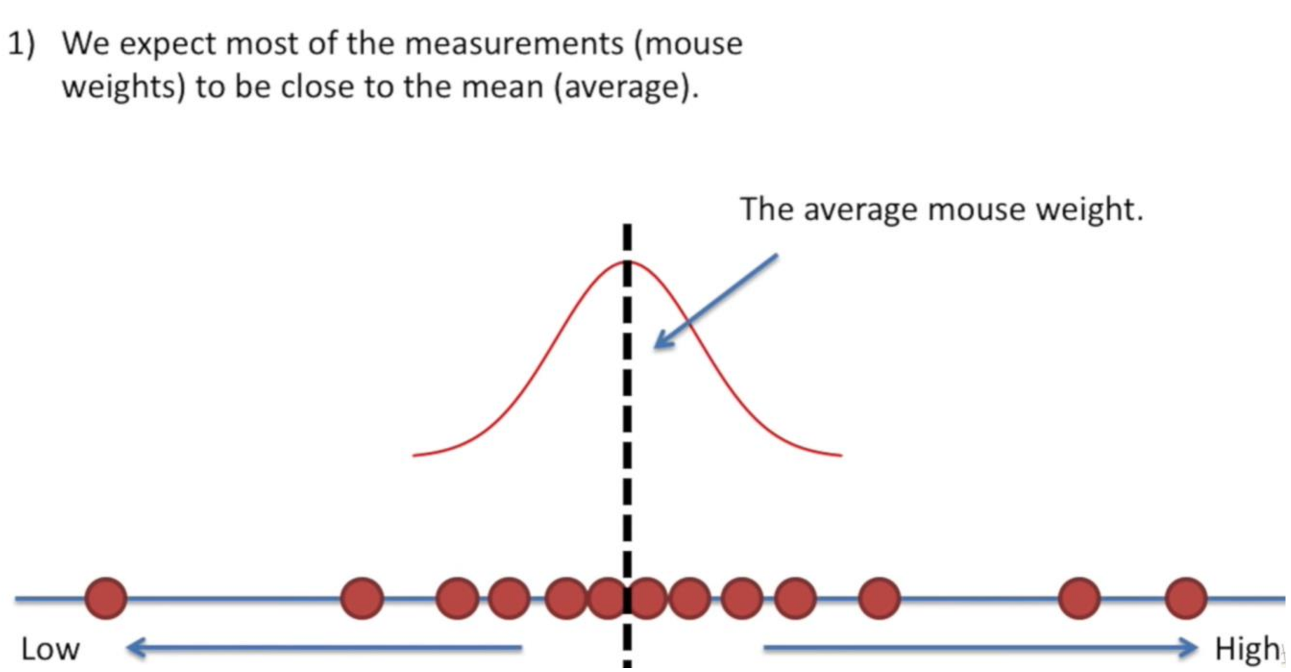
ML solution !

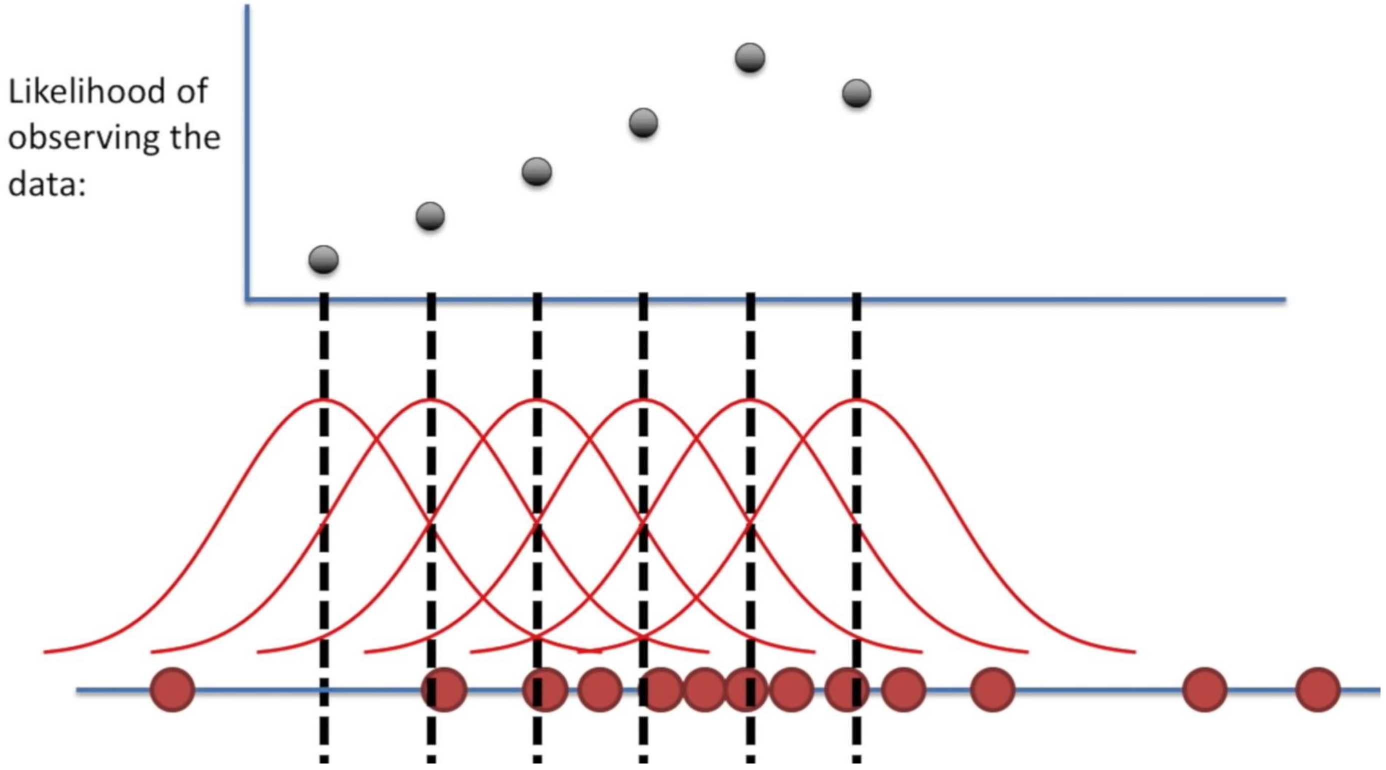
Maximum likelihood estimation

- The number of traffic accidents in Berkeley, California, in 10 randomly chosen nonrainy days in 1998 is as follows:
- 4, 0, 6, 5, 2, 1, 2, 0, 4, 3
- Use these data to estimate the proportion of nonrainy days that had 2 or fewer accidents that year.

MLE : Normal

- (Maximum Likelihood Estimator in a Normal Population) Suppose X1, . . . , Xn are independent, normal random variables each with unknown mean μ and unknown standard deviation σ.





MLE : Normal

- (Maximum Likelihood Estimator in a Normal Population) Suppose X1, . . . , Xn are independent, normal random variables each with unknown mean μ and unknown standard deviation σ.

$$f(x_1, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-(x_i - \mu)^2}{2\sigma^2}\right]$$

$$= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left[\frac{-\sum_1^n (x_i - \mu)^2}{2\sigma^2}\right]$$

$$\log f(x_1, \ldots, x_n | \mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{\sum_1^n (x_i - \mu)^2}{2\sigma^2}$$

MLE : Normal

- (Maximum Likelihood Estimator in a Normal Population) Suppose X1, . . . , Xn are independent, normal random variables each with unknown mean μ and unknown standard deviation σ.

$$\frac{\partial}{\partial \mu} \log f(x_1, \ldots, x_n | \mu, \sigma) = \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2}$$

$$\frac{\partial}{\partial \sigma} \log f(x_1, \ldots, x_n | \mu, \sigma) = -\frac{n}{\sigma} + \frac{\sum_{1}^{n} (x_i - \mu)^2}{\sigma^3}$$

$$\hat{\mu} = \sum_{i=1}^{n} x_i / n \qquad \hat{\sigma} = \left[\sum_{i=1}^{n} (x_i - \hat{\mu})^2 / n \right]^{1/2}$$

Model Selection

- Given some observations *X1, X2, …, XN*, how do you decide which probability distribution to model it ?


THANK YOU