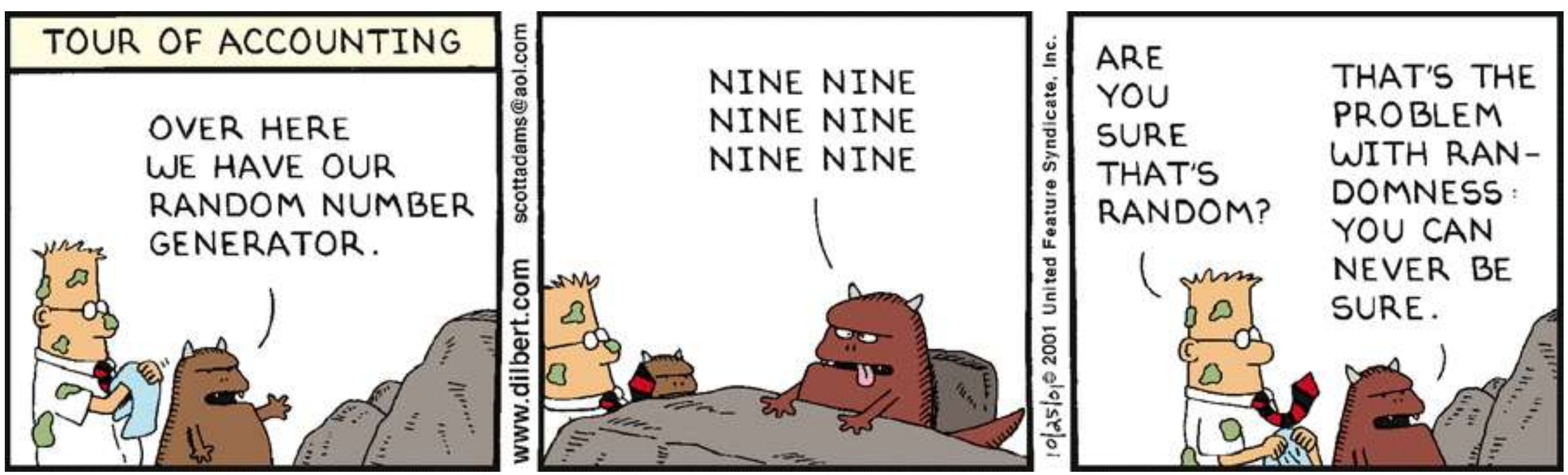
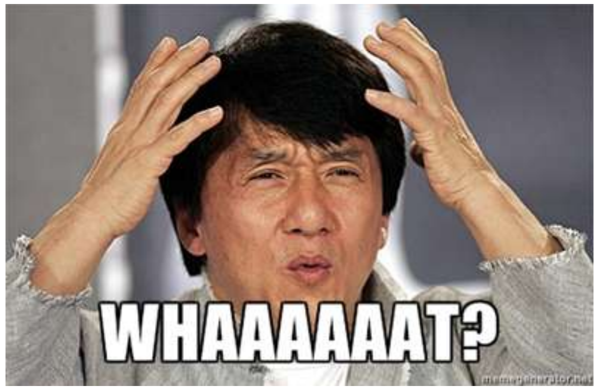
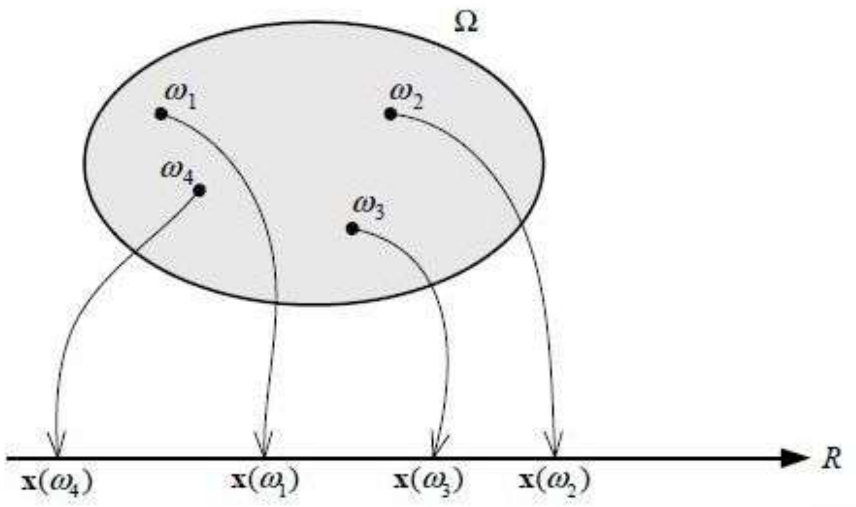
Random Variables



Random variables



- Rolling of two Dice
  - Sum is 7
  - sum is less than 3
- Random variable maps from Sample space to a real number $X : \Omega \rightarrow R$
- Probability of a random variable $P(X = 3) = P(\{\omega \in \Omega : X(\omega) = 3\})$



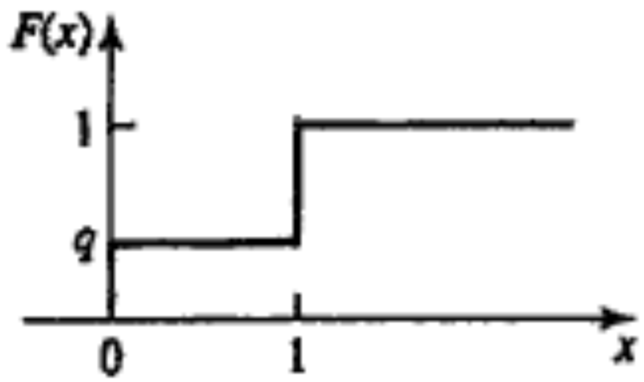
Example

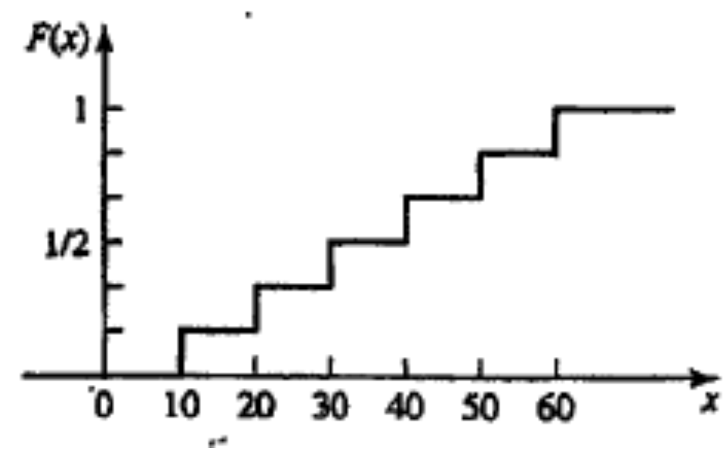
- Suppose that an individual purchases two electronic components, each of which may be either defective or acceptable.suppose that the four possible results — (d, d), (d, a), (a, d), (a, a) with probabilities .09, .21, .21, .49.
  - number of acceptable components obtained in the purchase
  - At least one acceptable component

Cumulative Distribution function

- F(x) = P(X <= x) = P($\{\omega \in \Omega : X(\omega) \leq x\}$)
- In the coin-tossing experiment, the probability of heads equals p and the probability of tails equals q. We define the random variable x such that X(h) = 1 X(t) = 0. Find the distribution function F(x)
- In the die experiment, we assign to the six outcomes the numbers X(i) = 10i.
  - Whats P(X < 35)
  - Plot F(x)

Cumulative distribution

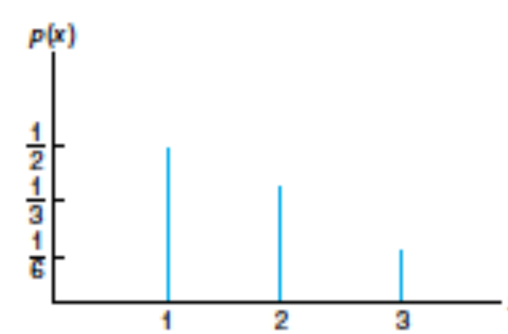


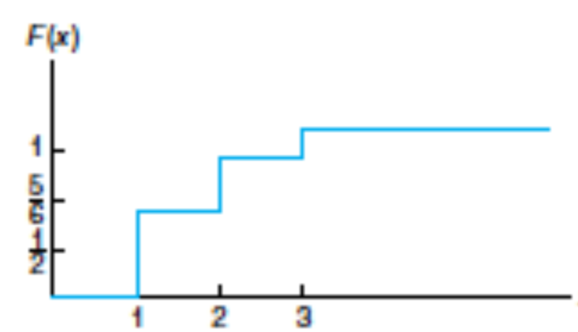


- All probability questions about X can be answered using F.
    - Find P{a < X ≤ b}.

Discrete random Variables

- Discrete RV
  - Possible values form a countable set which is either a finite set or a countably infinite set.
    - e.g. {0,1}, number of heads {0,…,N},
    - number of goals in a football match {0,1,…}
  - probability mass function P{X = a} = p(a)
    - p(x_i) >= 0, i = 1, 2, . . .
    - p(x) = 0, all other values of x
    - sum p(x_i) = 1

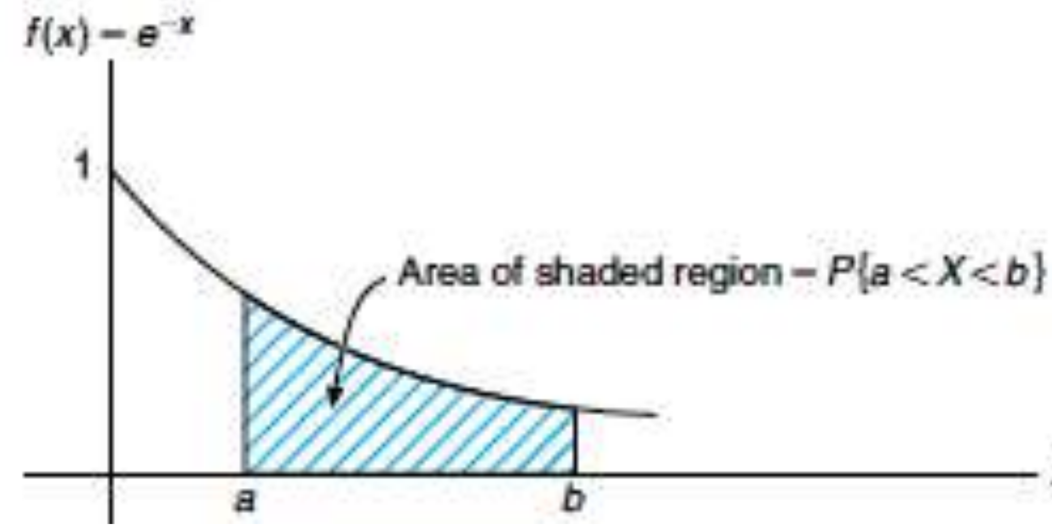




Continuous Random variables

- X takes values from a uncountable set
    - Time until next arrival [0, infty)
- Probability density function f(x)
- Probability that X = [a,b]

$$P\{a \leq X \leq b\} = \int_a^b f(x)\,dx$$



- Probability that X = a is 0!

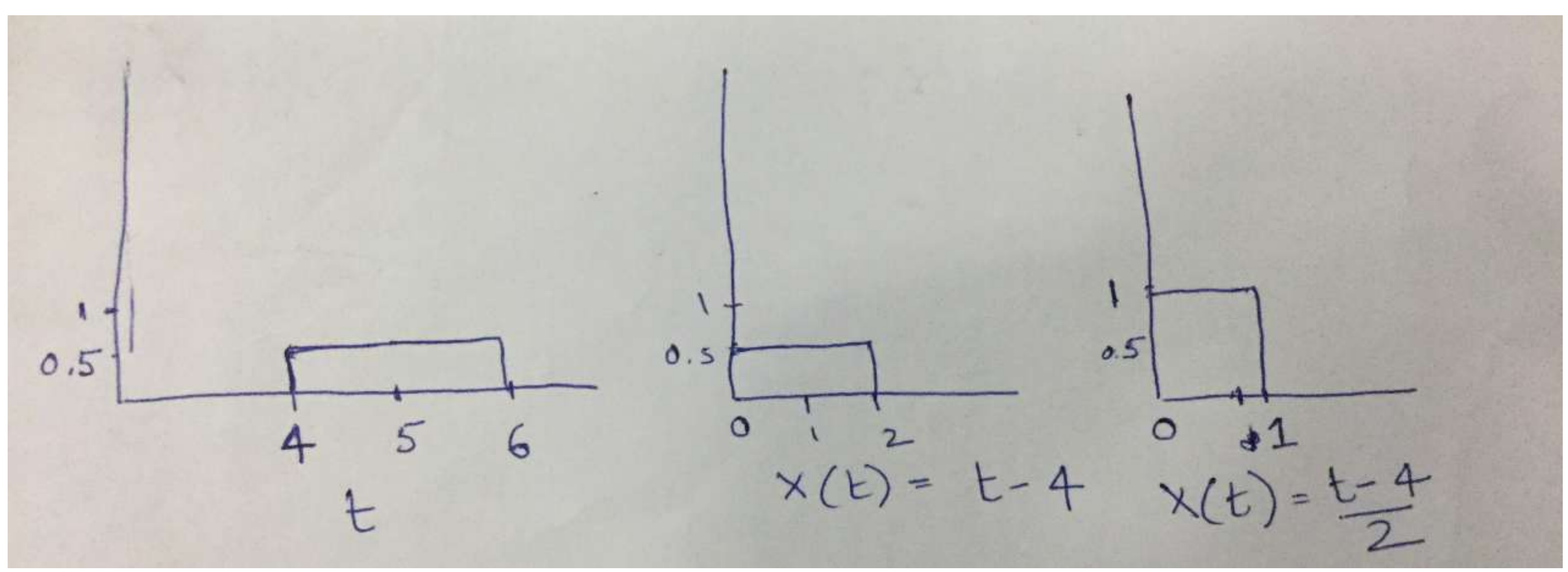
$$\int_a^a f(x)\,dx = 0 \qquad 1 = P\{X \in (-\infty, \infty)\} = \int_{-\infty}^{\infty} f(x)\,dx$$

- If not zero, probability sum to infinity
- CDF vs PDF

$$\frac{d}{da}F(a) = f(a) \qquad F(a) = P\{X \in (-\infty, a]\} = \int_{-\infty}^{a} f(x)\,dx$$

Continuous R.V.

- Suppose a species of bacteria typically lives 4 to 6 hours. What is the probability that a bacterium lives exactly 5 hours?
- What is the probability that the bacterium dies between 5 hours and 5.1 hours?
- probability that the bacterium dies within a small (infinitesimal) window of time around 5 hours : 0.5 dt
- Probability Density function : f(x) dx as being the probability of X falling within the infinitesimal interval [x, x + dx].



Example

- Suppose that X is a continuous random variable whose probability density function is given by

$$f(x) = \begin{cases} C(4x - 2x^2) & 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

  - (a) What is the value of C?
  - (b) Find $P\{X > 1\}$.

Expectation

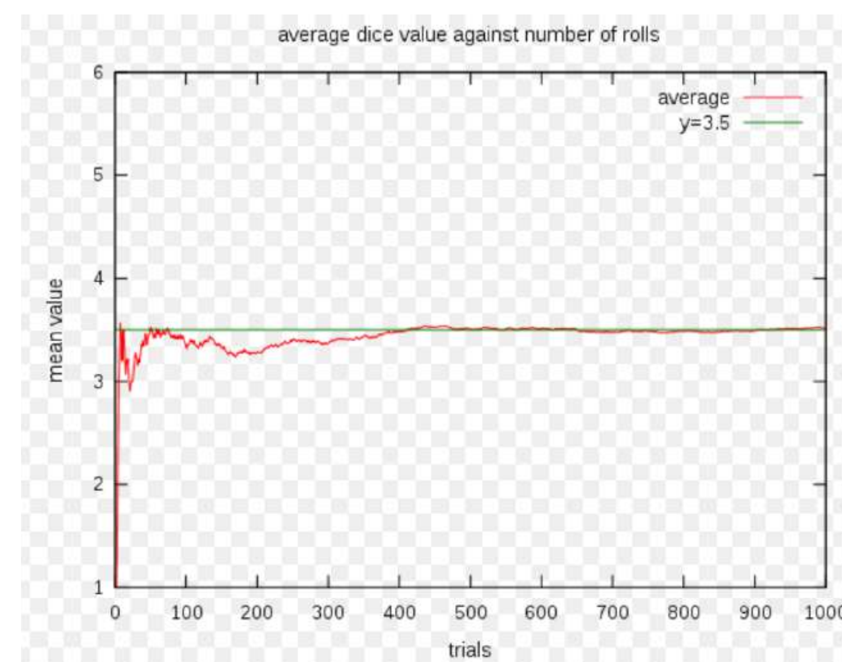
- Expected value of a random variable is the long-run average value of repetitions of the experiment

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

- Discrete random variable is the probability-weighted average of all possible values.
  - Rolling a fair sided dice

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

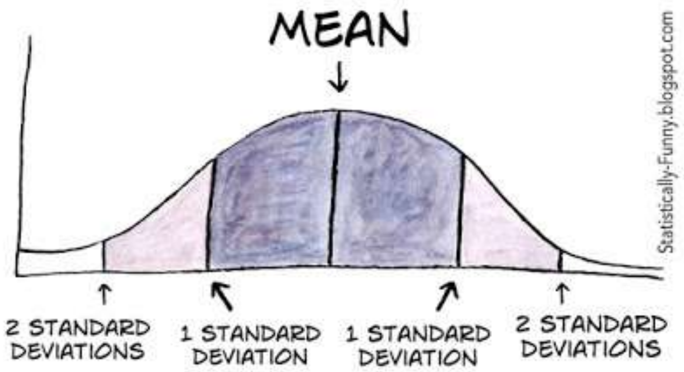
- Continous r.v.
  - Prove $E[aX + b] = aE[X] + b$



Expectation

- Prove $E[aX + b] = aE[X] + b$
- Suppose that you are expecting a message at some time past 5 P.M. From experience you know that X , the number of hours after 5 P.M. until the message arrives, is a random variable with the following probability density function: Whats expected amount of time past 5 P.M. until the message arrives ?

$$f(x) = \begin{cases} \frac{1}{1.5} & \text{if } 0 < x < 1.5 \\ 0 & \text{otherwise} \end{cases}$$

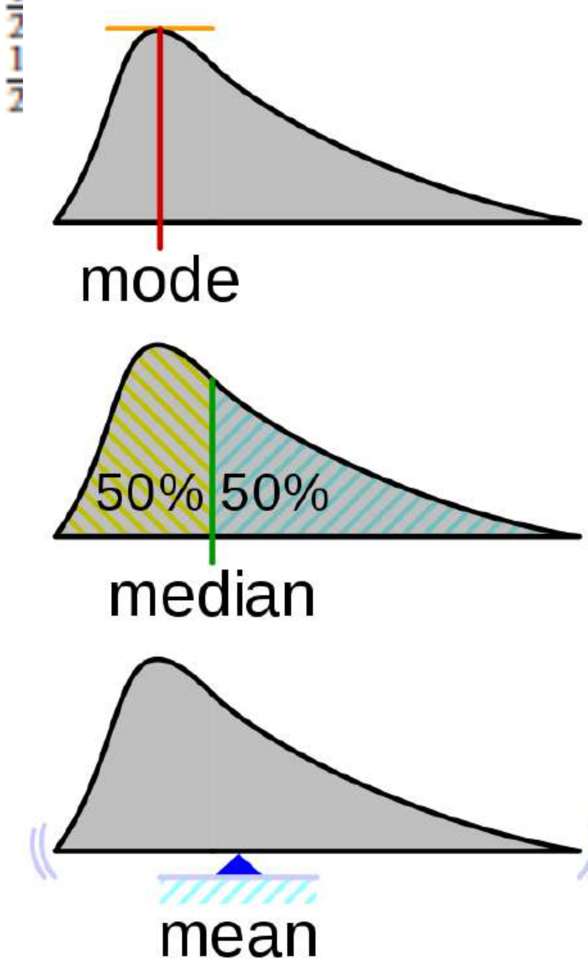


Variance

SIX HOURS AFTER LEAVING FOR A SHORT DRIVE, GREGORY HAD TO ADMIT HE HAD PROBABLY TAKEN A WRONG TURN.

- Spread of the random variable values

$W = 0 \quad \text{with probability } 1 \qquad Y = \begin{cases} -1 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases} \qquad Z = \begin{cases} -100 & \text{with probability } \frac{1}{2} \\ 100 & \text{with probability } \frac{1}{2} \end{cases}$



- Variance : $\mathrm{Var}(X) = E[(X-\mu)^2] \quad = E[X^2] - (E[X])^2$
    - Variance of fair sided die
    - Prove $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$

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

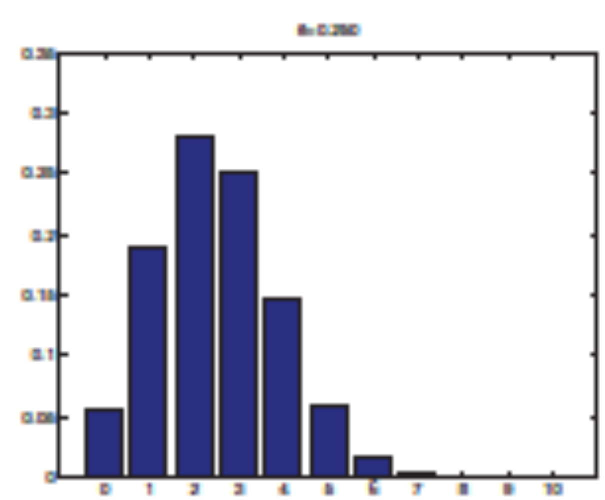
Common Discrete Distributions

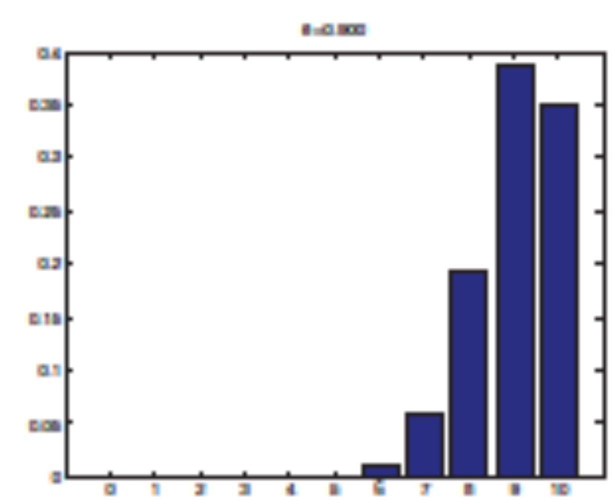
- Let $X \in \{0, 1\}$ be a binary random variable, with probability of "success" θ, X has a Bernoulli distribution, $X \sim \mathrm{Ber}(\theta)$
  - Coin toss, Rain or not $\mathrm{Ber}(x|\theta) = \theta^{\mathbb{I}(x=1)}(1-\theta)^{\mathbb{I}(x=0)}$ $\mathrm{Ber}(x|\theta) = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$
- 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 $X \sim \mathrm{Bin}(n, \theta)$.

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

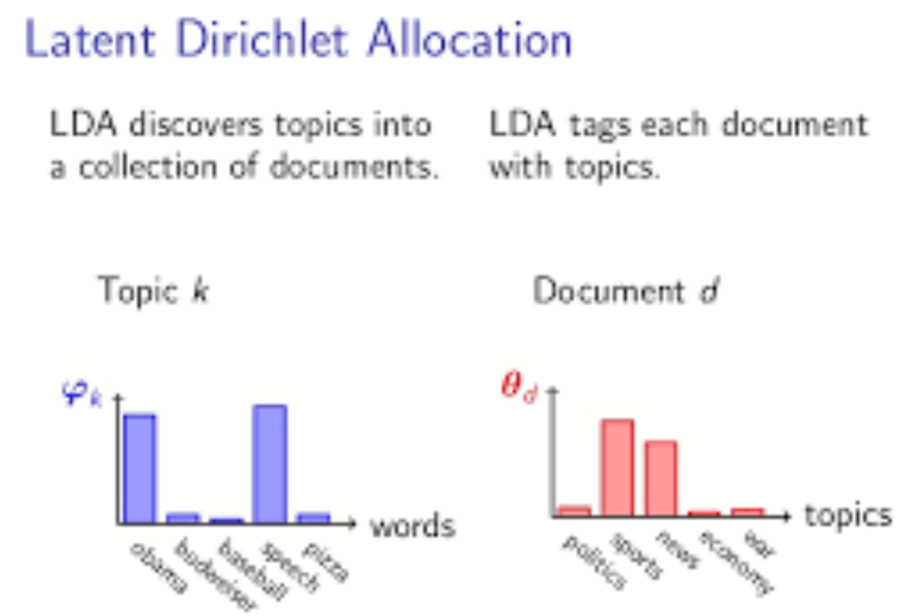
$$\text{mean} = n\,\theta, \quad \mathrm{var} = n\theta(1-\theta)$$

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





Discrete Distributions

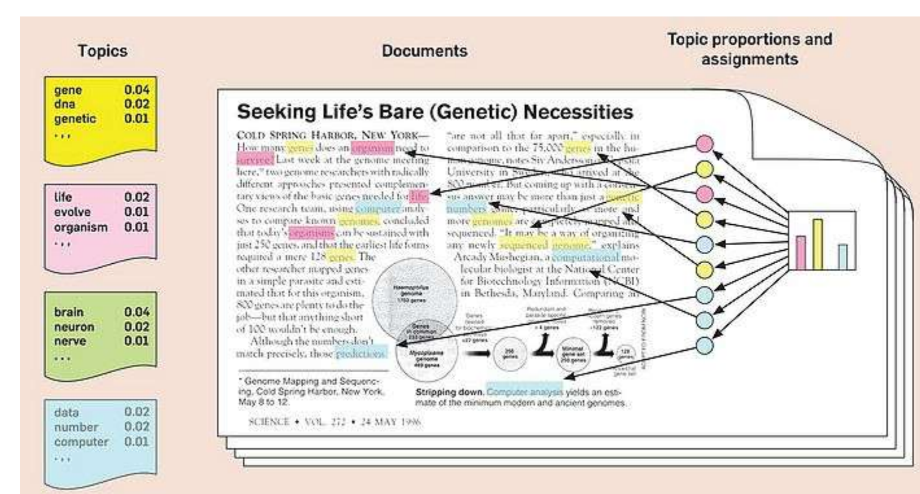


- Model the outcomes of tossing a K -sided die : categorical/Multinoulli distribution, x ~ Cat(θ), p(x = j|θ) = θj .
- Multinomial distribution : Models the outcome of n dice rolls, let x = (x1, . . . , xK) be a random vector, where xj number of times side j of the die occurs.

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

- Probabilistic topic model
- Text classification

$$\binom{n}{x_1 \ldots x_K} \triangleq \frac{n!}{x_1! x_2! \cdots x_K!}$$



Text Modelling

- Document :
- Vocabulary :
- Representation :
- Bag of Words :

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow

mary	lamb	little	big	fleece	white	black	snow	rain	unk
1	2	3	4	5	6	7	8	9	10

1 10 3 2 3 2 3 2
1 10 3 2 10 5 10 6 8

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	1	1	1	0	1	1	0	1	0	1

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

a t a g c c g g t a c g g c a
t t a g c t g c a a c c g c a
t c a g c c a c t a g a g c a
a t a a c c g c g a c c g c a
t t a g c c g c t a a g g t a
t a a g c c t c g t a c g t a
t t a g c c g t t a c g g c c
a t a t c c g g t a c a g t a
a t a g c a g g t a c c g a a
a c a t c c g t g a c g g a a

class 1	prob	class 2	prob
subject	0.998	subject	0.998
this	0.628	windows	0.639
with	0.535	this	0.540
but	0.471	with	0.538
you	0.431	but	0.518

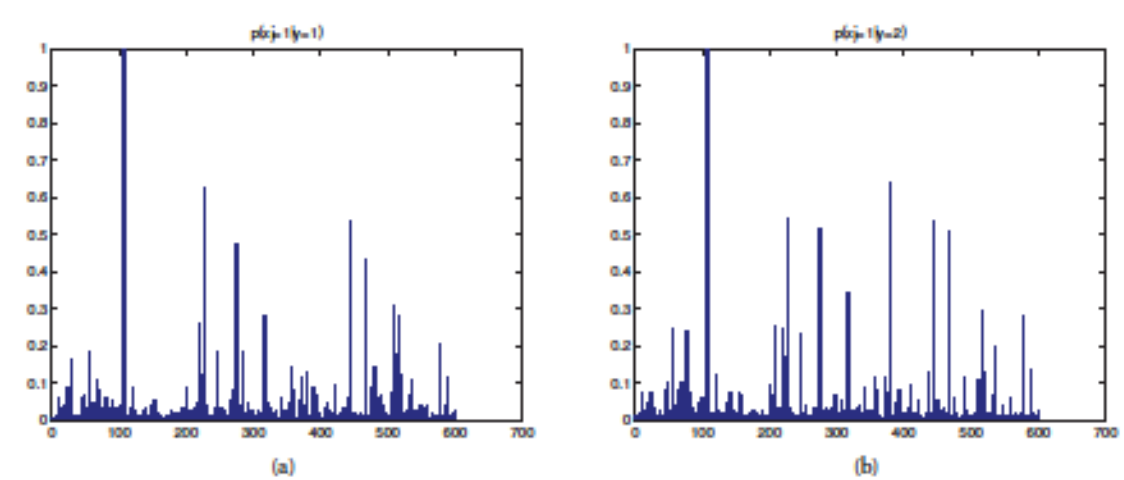
- xi be a vector of counts for document i, θjc is the probability of generating word j in documents of class c;

- Na $p(\mathbf{x}_i|y_i = c, \boldsymbol{\theta}) = \mathrm{Mu}(\mathbf{x}_i|N_i, \boldsymbol{\theta}_c) = \frac{N_i!}{\prod_{j=1}^{D} x_{ij}!} \prod_{j=1}^{D} \theta_{jc}^{x_{ij}}$

-

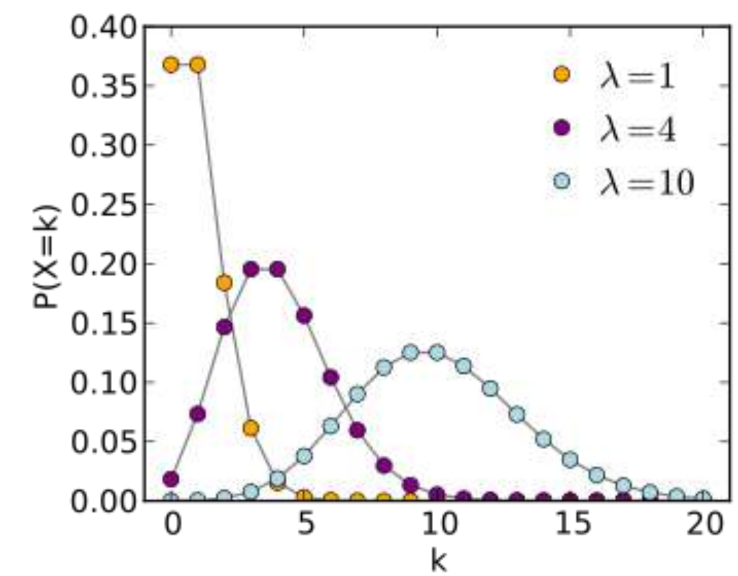
$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} \mathrm{Ber}(x_j|\mu_{jc})$$

$$p(y = c|\mathbf{x}, \mathcal{D}) \propto p(y = c|\mathcal{D}) \prod_{j=1}^{D} p(x_j|y = c, \mathcal{D})$$



Poisson distribution

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



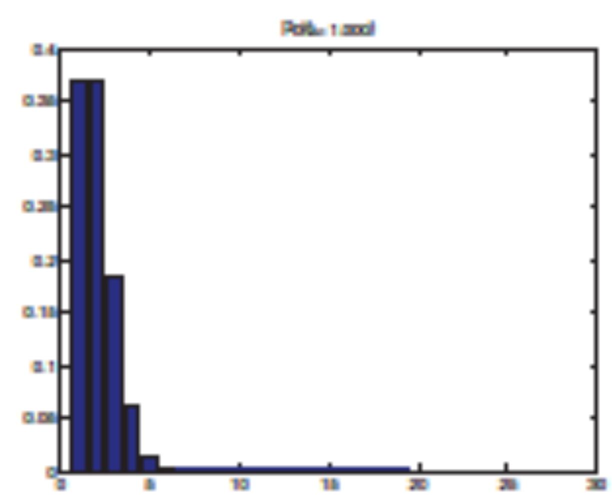
- Model number of events occurring in a fixed interval of time/space

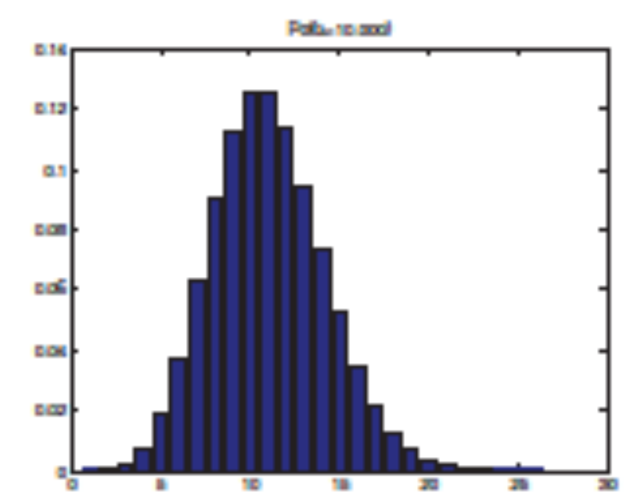
$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$

- 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 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.





Poisson distribution

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

- Modeling rare events : Approximation for a binomial r.v when n is large and p is small, λ = np.

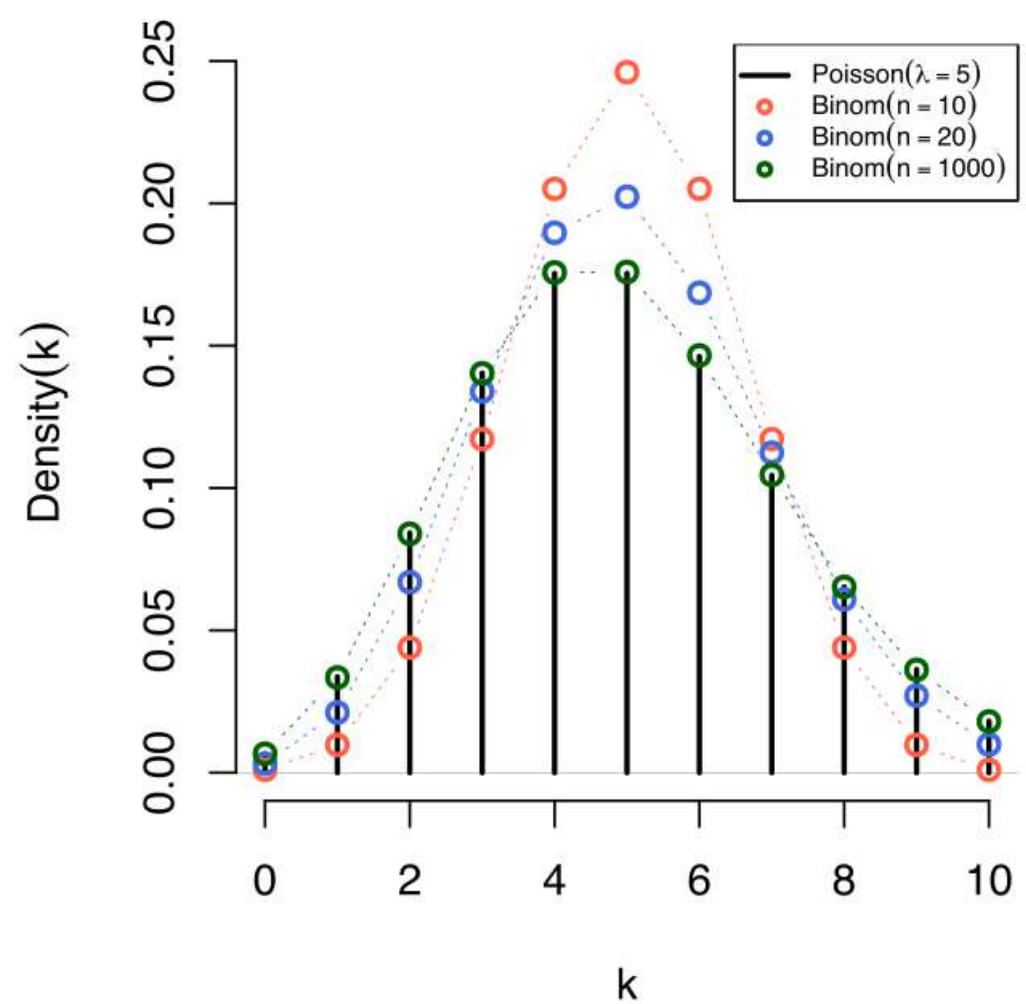
$$P\{X=i\} = \frac{n!}{(n-1)!i!}p^i(1-p)^{n-i}$$
$$= \frac{n!}{(n-1)!i!}\left(\frac{\lambda}{n}\right)^i\left(1-\frac{\lambda}{n}\right)^{n-i}$$
$$= \frac{n(n-1)\dots(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}$$

Now, for n large and p small,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda} \quad \frac{n(n-1)\dots(n-i+1)}{n^i} \approx 1 \quad \left(1-\frac{\lambda}{n}\right)^i \approx 1$$

Hence, for n large and p small,

$$P\{X=i\} \approx e^{-\lambda}\frac{\lambda^i}{i!}$$



- Poisson distribution violations
- The number of emails you receive in a day
- Number of high magnitude earthquakes

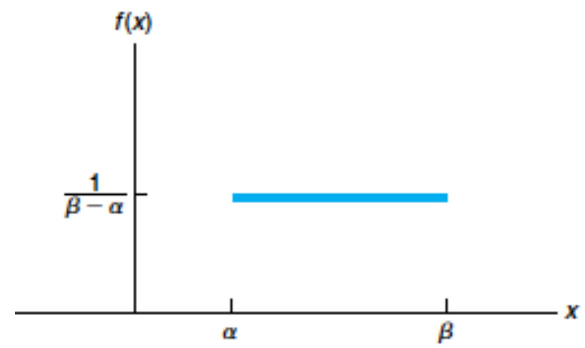
Examples

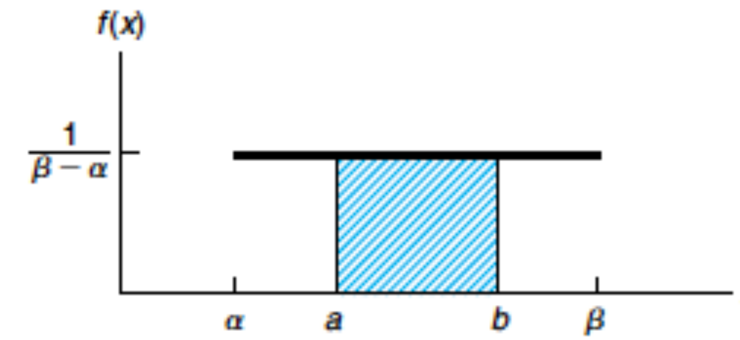
- It is known that disks produced by a certain company will be defective with probability .01 independently of each other. The company sells the disks in packages of 10 and offers a money-back guarantee if more than 1 of the 10 disks is defective. What proportion of packages is returned?
    - Using Binomial distribution assumption
    - Using Poisson distribution assumption

Uniform Random Variables

- Uniform random variable : X is said to be uniformly distributed over the interval [α, β]

$$f(x) = \begin{cases} \frac{1}{\beta - \alpha} & \text{if } \alpha \le x \le \beta \\ 0 & \text{otherwise} \end{cases}$$



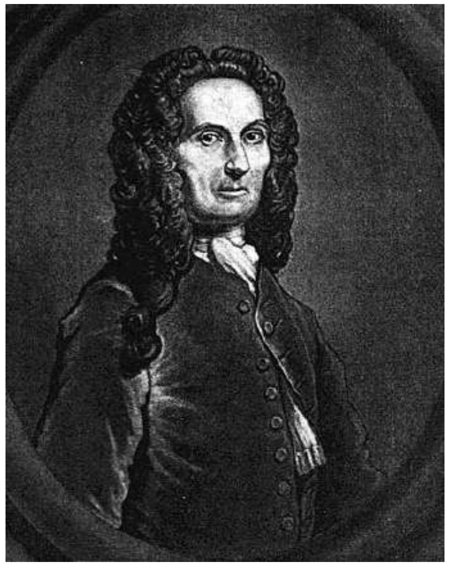


- Probability that X lies in [a,b]

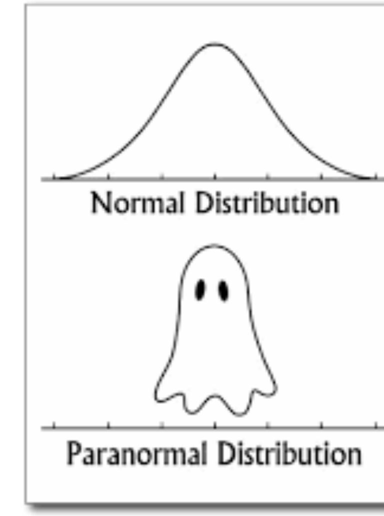
$$P\{a < X < b\} = \frac{1}{\beta - \alpha} \int_a^b dx \quad = \frac{b-a}{\beta - \alpha}$$

$$E[X] = \frac{\alpha + \beta}{2} \qquad \text{Var}(X) = \frac{(\beta - \alpha)^2}{12}$$

- Buses arrive at a specified stop at 15-minute intervals starting at 7 A.M. That is, they arrive at 7, 7:15, 7:30, 7:45, and so on. If a passenger arrives at the stop at a time that is uniformly distributed between 7 and 7:30, find the probability that he waits
- (a) less than 5 minutes for a bus;
- (b) at least 12 minutes for a bus.

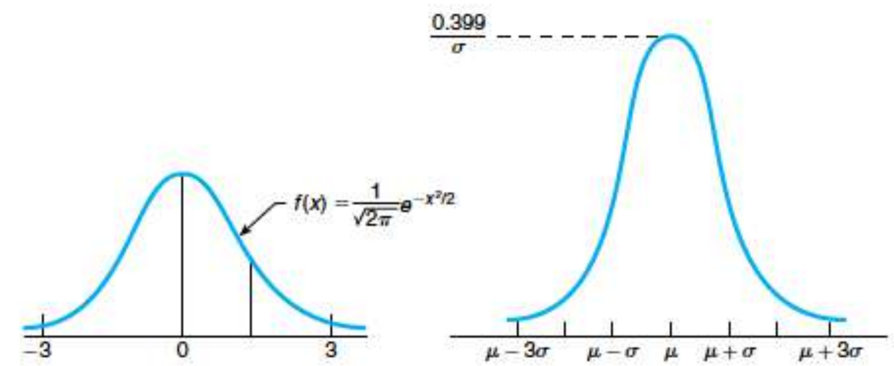


Normal Random Variables



- 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 parameters μ and σ2, X ~ N(μ, σ2)

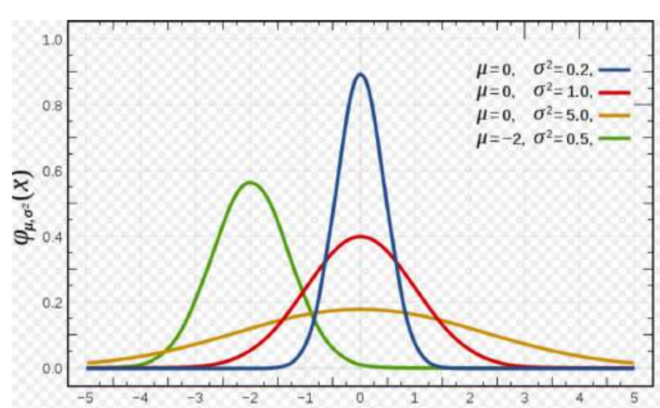
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty^*$$

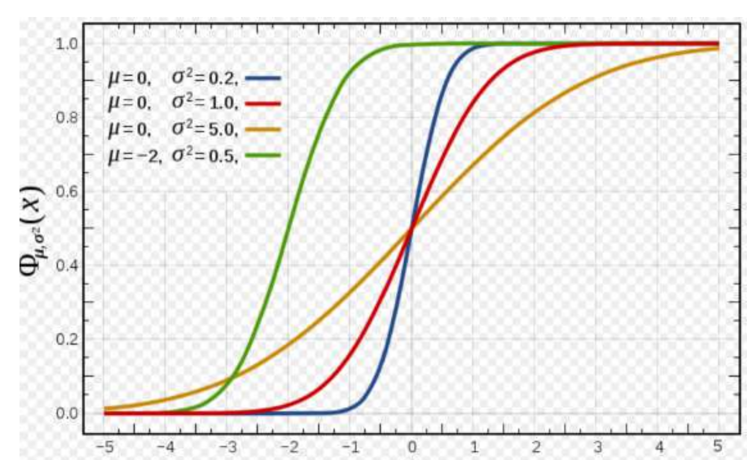


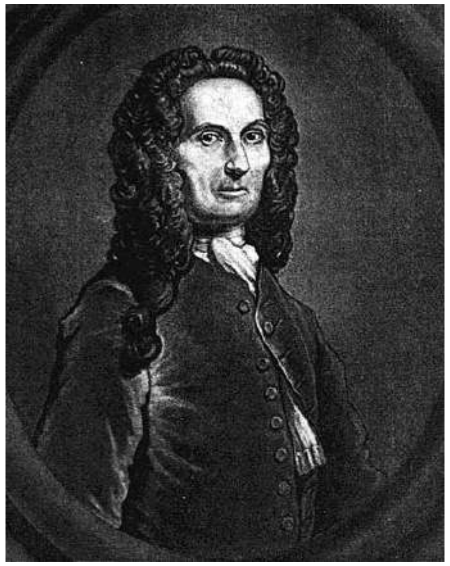
- μ = E [X] is the mean (and mode), and σ2 = var [X] is the variance.

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

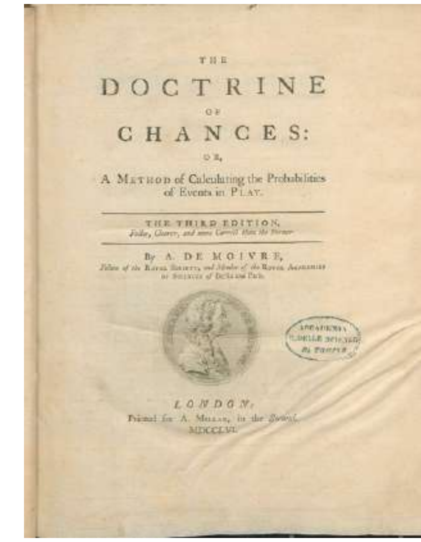
- CDF of the Gaussian







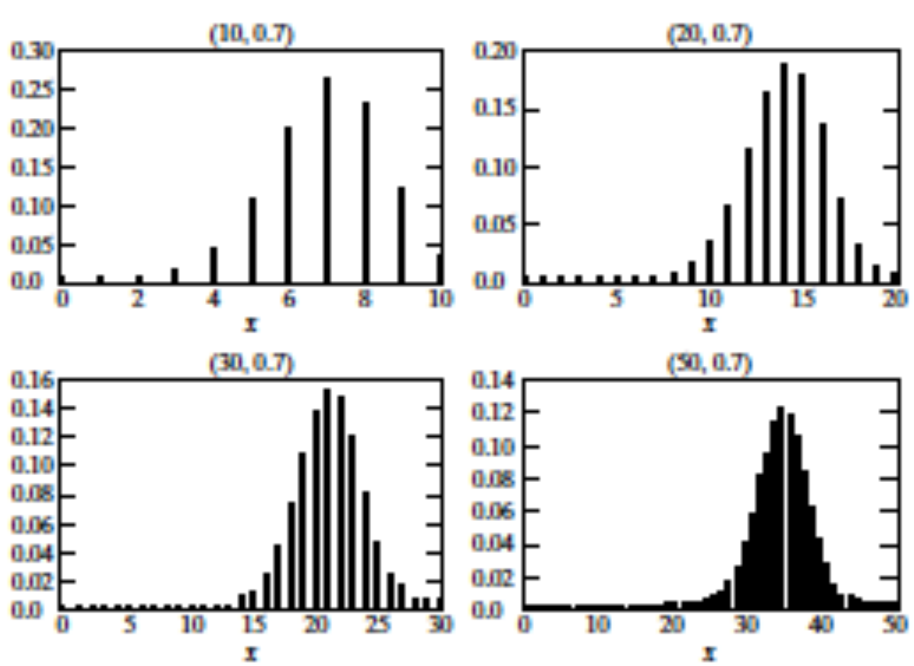
Normal Random Variables

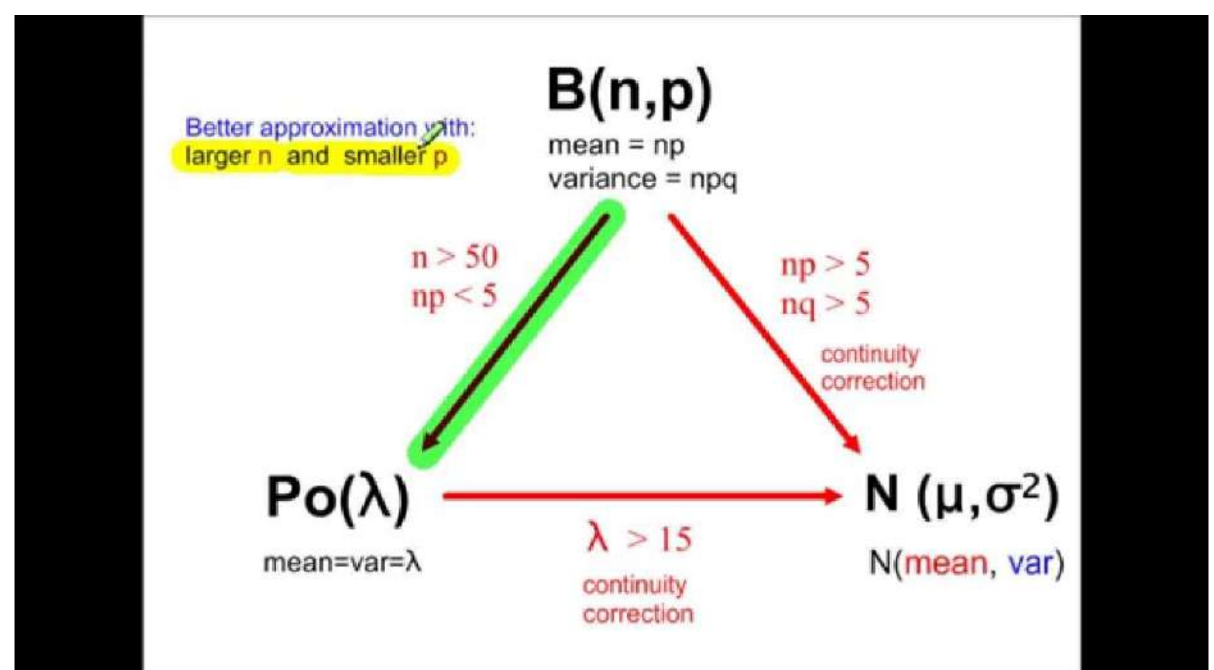


- A random variable is said to be normally distributed with parameters μ and σ2, X ~ N(μ, σ2)

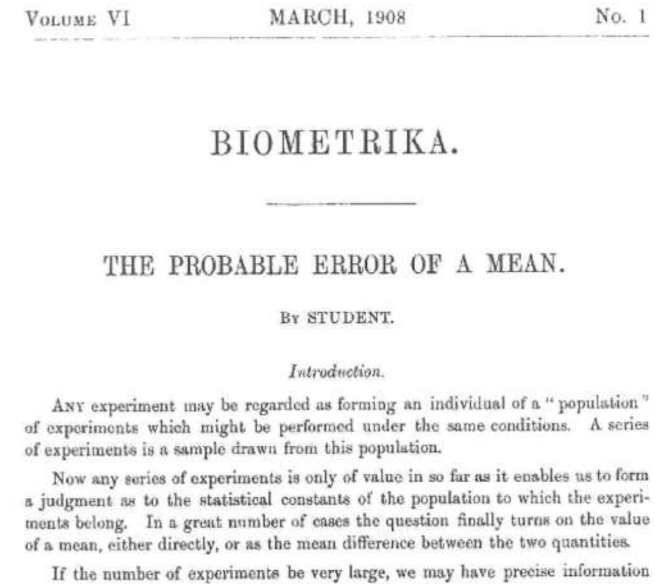
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}, \qquad -\infty < x < \infty^*$$

- Approximates Binomial for large n; calculate #heads >60 in 100 tosses.
- Central limit Theorem (Laplace, 1778) : the means of repeated samples from the distribution (not normal) will be normally distributed





Student's t-distribution


VOLUME VI MARCH, 1908 No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

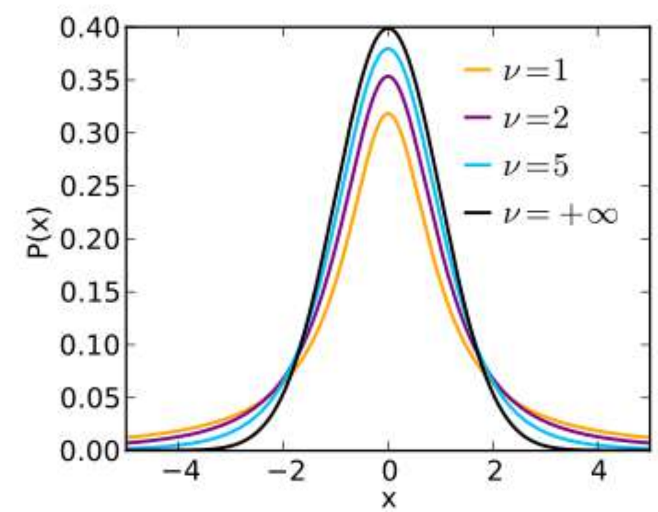
Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information

- heavy tailed distribution
- Student t-distribution
- Laplace distribution

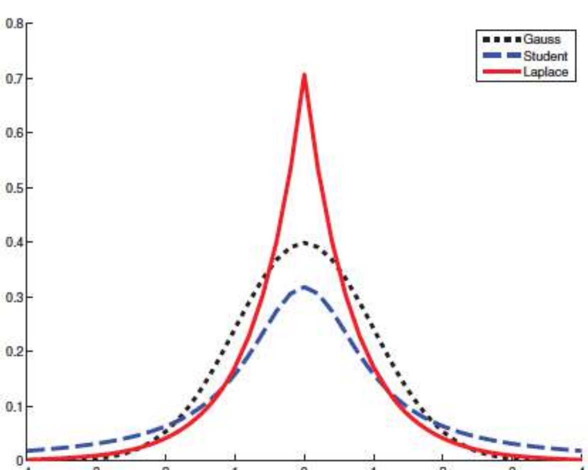
$$\mathcal{T}(x|\mu,\sigma^2,\nu) \propto \left[1+\frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\left(\frac{\nu+1}{2}\right)}$$

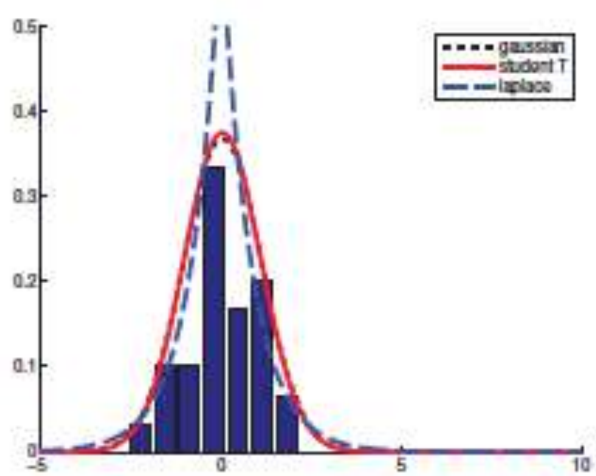
$$\text{mean} = \mu, \text{mode} = \mu, \ \text{var} = \frac{\nu\sigma^2}{(\nu-2)}$$

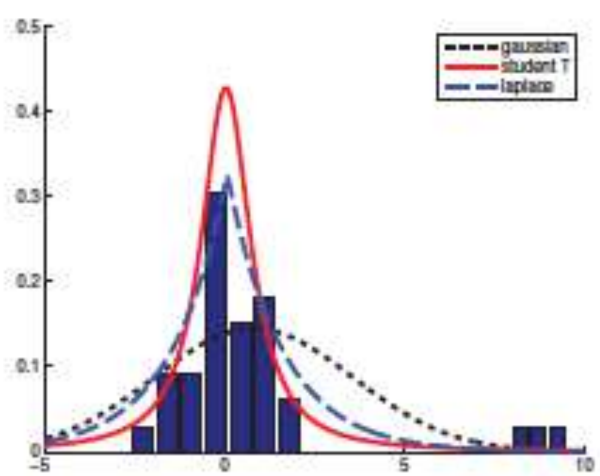


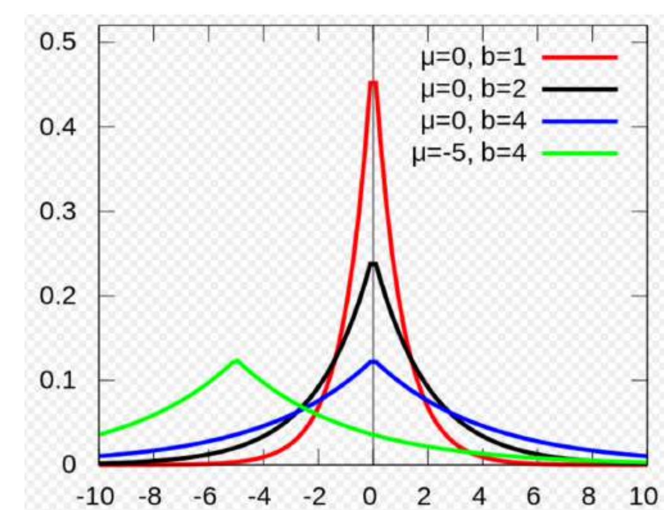
$$\text{Lap}(x|\mu,b) \triangleq \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

$$\text{mean} = \mu, \ \text{mode} = \mu, \ \text{var} = 2b^2$$









Student t-test

- if mean of a population has a value specified in a null hypothesis.
- check if means of two populations are equal
- slope of a regression line differs significantly from 0.

Subject #	Score 1	Score 2	X-Y	(X-Y)^2
1	3	20	-17	289
2	3	13	-10	100
3	3	13	-10	100
4	12	20	-8	64
5	15	29	-14	196
6	16	32	-16	256
7	17	23	-6	36
8	19	20	-1	1
9	23	25	-2	4
10	24	15	9	81
11	32	30	2	4
		SUM:	**-73**	**1131**

Microsoft Excel 2010 and later	`T.TEST(array1, array2, tails, type)`
LibreOffice	`TTEST(Data1; Data2; Mode; Type)`
Google Sheets	`TTEST(range1, range2, tails, type)`
Python	`scipy.stats.ttest_ind(a, b, axis=0, equal_var=True)`
Matlab	`ttest(data1, data2)`

$$t = \frac{(\Sigma D)/N}{\sqrt{\frac{\Sigma D^2 - \left(\frac{(\Sigma D)^2}{N}\right)}{(N-1)(N)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{(-73)^2}{11}\right)}{(11-1)(11)}}}$$

$$t = \frac{-73/11}{\sqrt{\frac{1131 - \left(\frac{5329}{11}\right)}{110}}}$$

$$t = -\ 2.74$$

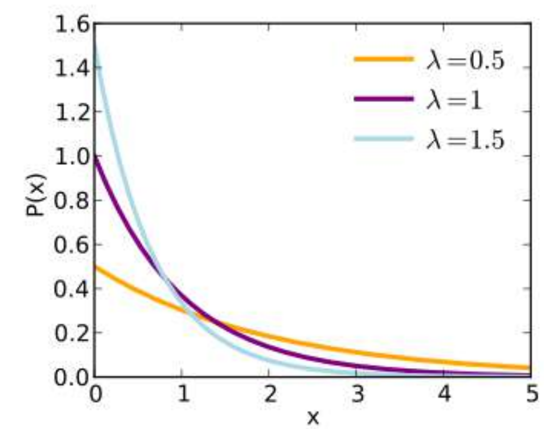
Two Tails T Distribution Table

DF	A = 0.2	0.10	0.05	0.02
∞	t_a = 1.282	1.645	1.960	2.326
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	3.365
6	1.440	1.943	2.447	3.143
7	1.415	1.895	2.365	2.998
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764

Exponential Random variables

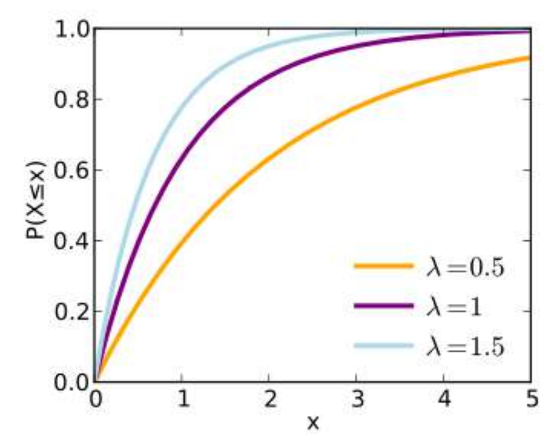
- Distribution of the amount of time until some specific event occurs.
  - the amount of time until an earthquake occurs, a new war breaks out
- X is exponentially distributed with rate parameter λ > 0

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \qquad F(x) = P\{X \leq x\} \quad 1 - e^{-\lambda x},$$



- Exponential random variable is memoryless, distribution of additional functional life of an item of age t is the same as that of a new item

$$P\{X > s + t | X > t\} = P\{X > s\} \qquad P\{X > s + t\} = P\{X > s\}P\{X > t\}$$



- Suppose that a number of miles that a car can run before its battery wears out is exponentially distributed with an average value of 10,000 miles. If a person desires to take a 5,000-mile trip, what is the probability that she will be able to complete her trip without having to replace her car battery?

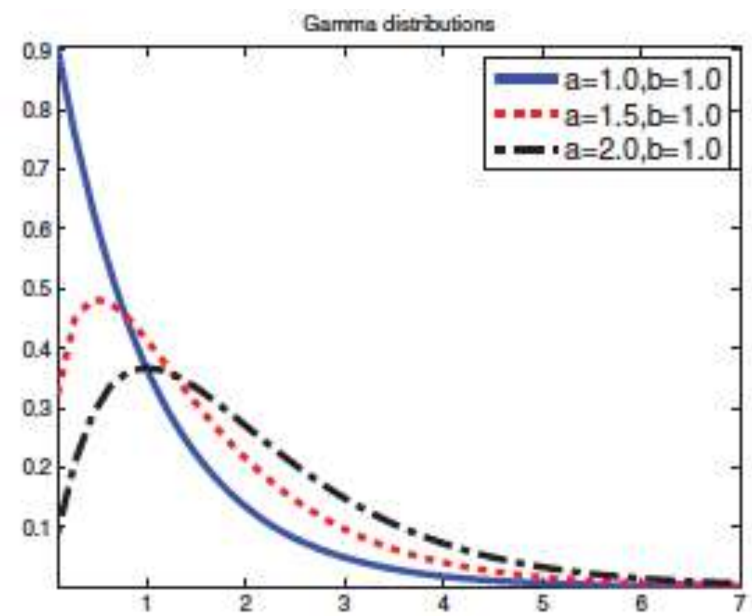
Gamma and Beta Distributions

- Gamma distribution for positive real valued rv's, x > 0, is defined in terms of two parameters, shape a > , rate b > 0:

$$\mathrm{Ga}(T|\mathrm{shape}=a, \mathrm{rate}=b) \triangleq \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb}$$

$$\mathrm{mean} = \frac{a}{b}, \ \mathrm{mode} = \frac{a-1}{b}, \ \mathrm{var} = \frac{a}{b^2}$$

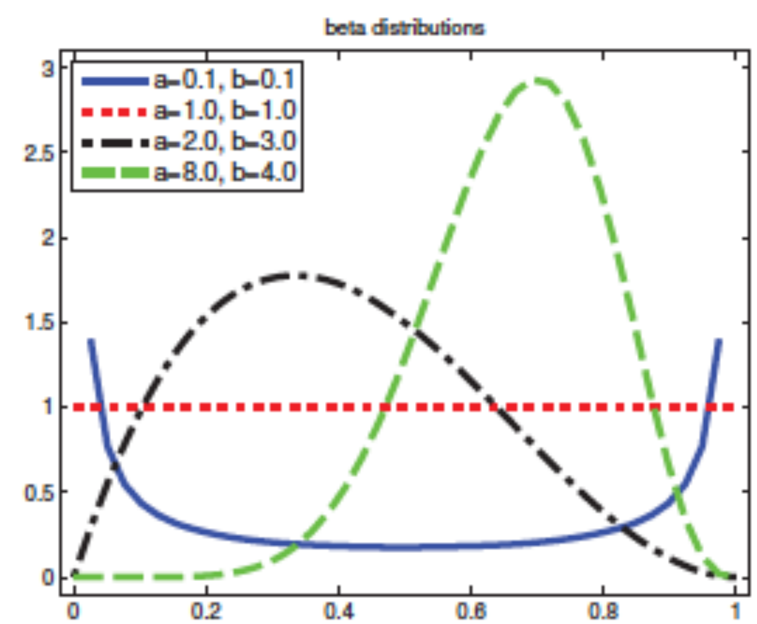
$$\Gamma(x) \triangleq \int_0^\infty u^{x-1} e^{-u} du$$



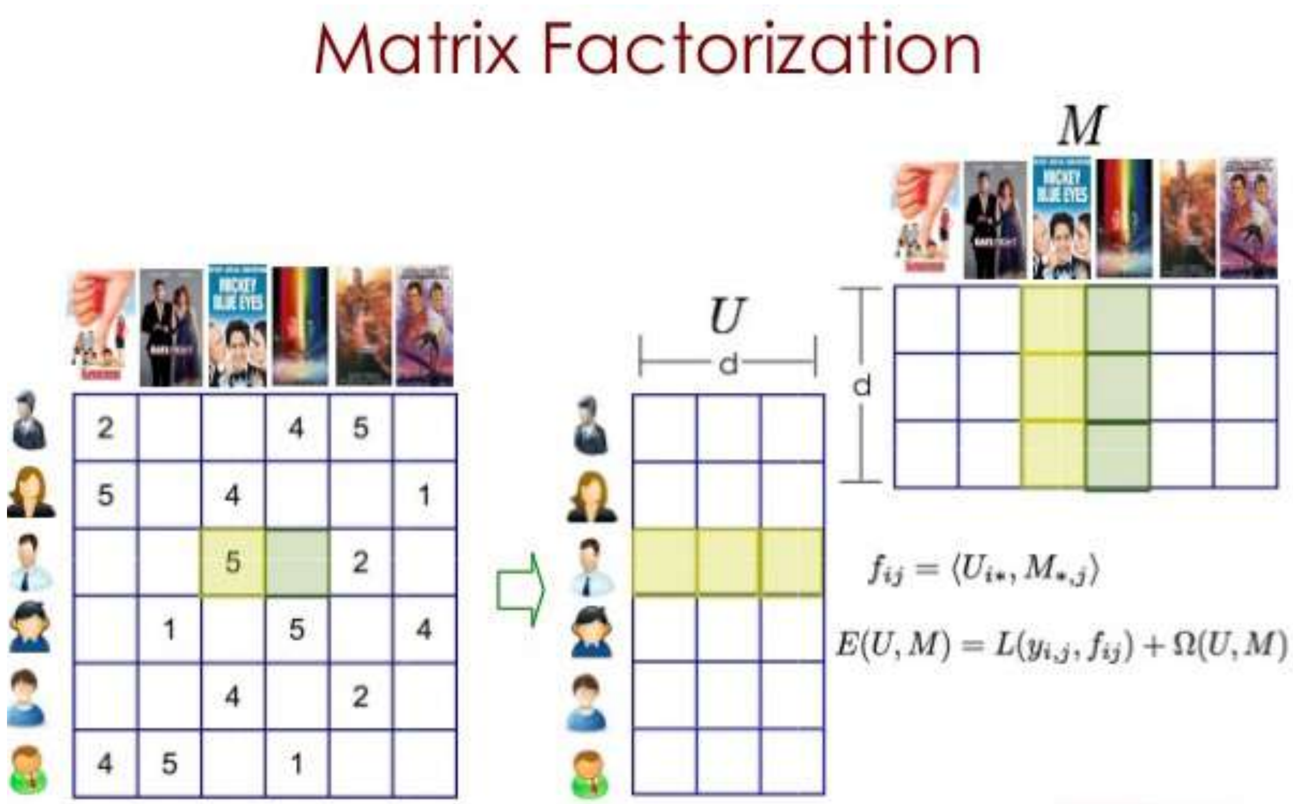
- Exponential: Expon(x|λ) = Ga(x| 1, λ), sum of n independent exponential r.v. is a Gamma r.v. Ga(x|n, λ)
  - a stereo cassette requires one battery to operate, then the total playing time one can obtain from a total of n batteries
- Beta distribution has support in the interval [0, 1]
  - model events which are constrained to take place within an interval with a minimum and maximum value : project management
- Useful as prior in Bayesian modelling

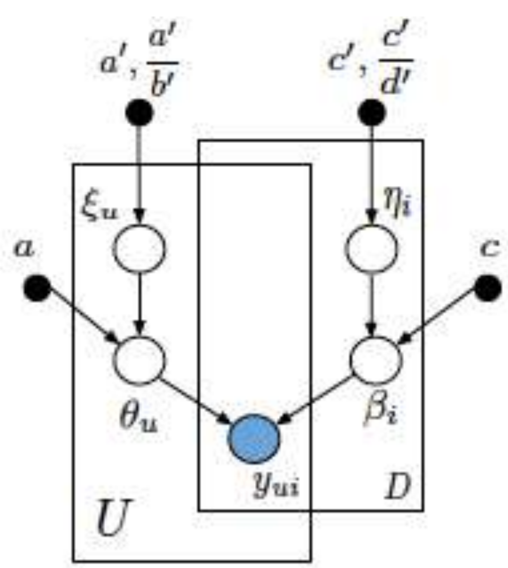
$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \quad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\mathrm{mean} = \frac{a}{a+b}, \ \mathrm{mode} = \frac{a-1}{a+b-2}, \ \mathrm{var} = \frac{ab}{(a+b)^2(a+b+1)}$$



Poisson Matrix Factorization





1. For each user u:
   (a) Sample activity $\xi_u \sim \text{Gamma}(a', a'/b')$.
   (b) For each component k, sample preference

$$\theta_{uk} \sim \text{Gamma}(a, \xi_u).$$

2. For each item i:
   (a) Sample popularity $\eta_i \sim \text{Gamma}(c', c'/d')$.
   (b) For each component k, sample attribute

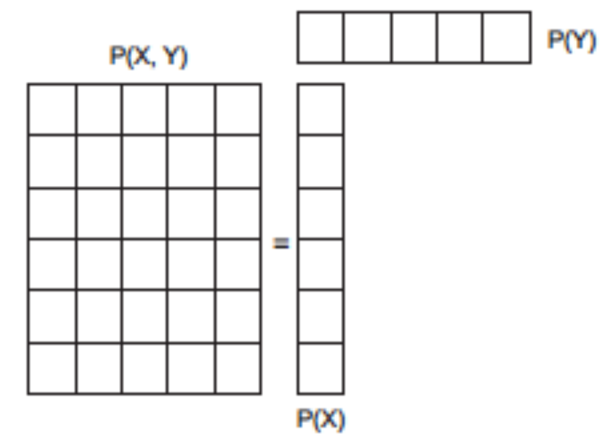
$$\beta_{ik} \sim \text{Gamma}(c, \eta_i).$$

3. For each user u and item i, sample rating

$$y_{ui} \sim \text{Poisson}(\theta_u^\top \beta_i).$$

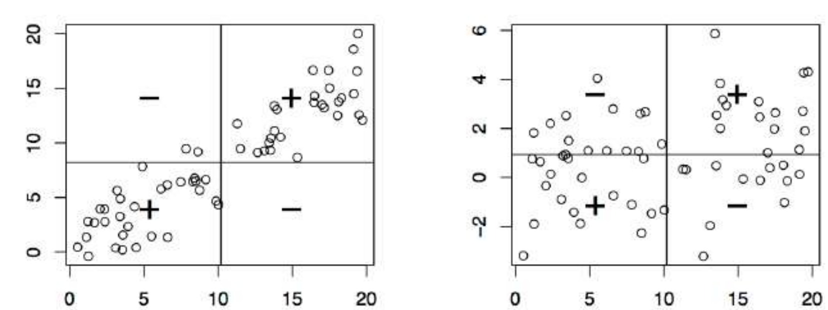
Joint Probability Distributions

- p(x1 , . . . , xD) : models the (stochastic) relationships between the variables.
- discrete variables : multi-dimensional array, number of parameters is O(K^D)
- Covariance between measures the degree to which X and Y are (linearly) related.

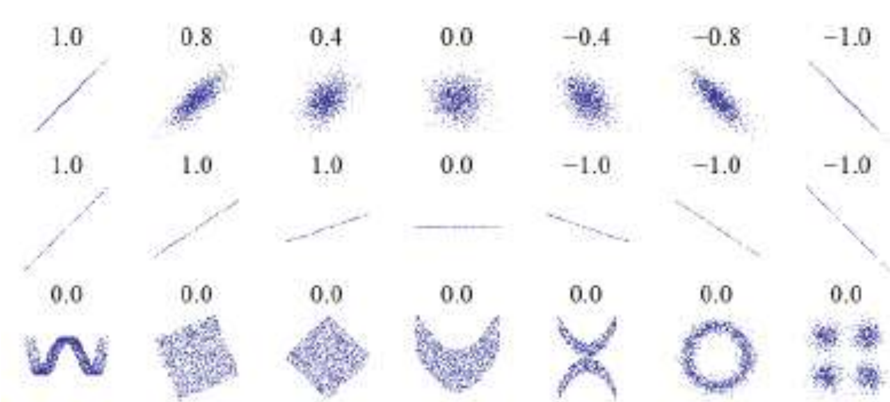


$$X \perp Y \iff p(X,Y) = p(X)p(Y)$$

$$\mathrm{cov}\,[X,Y] \triangleq \mathbb{E}\,[(X-\mathbb{E}\,[X])(Y-\mathbb{E}\,[Y])] = \mathbb{E}\,[XY] - \mathbb{E}\,[X]\,\mathbb{E}\,[Y]$$

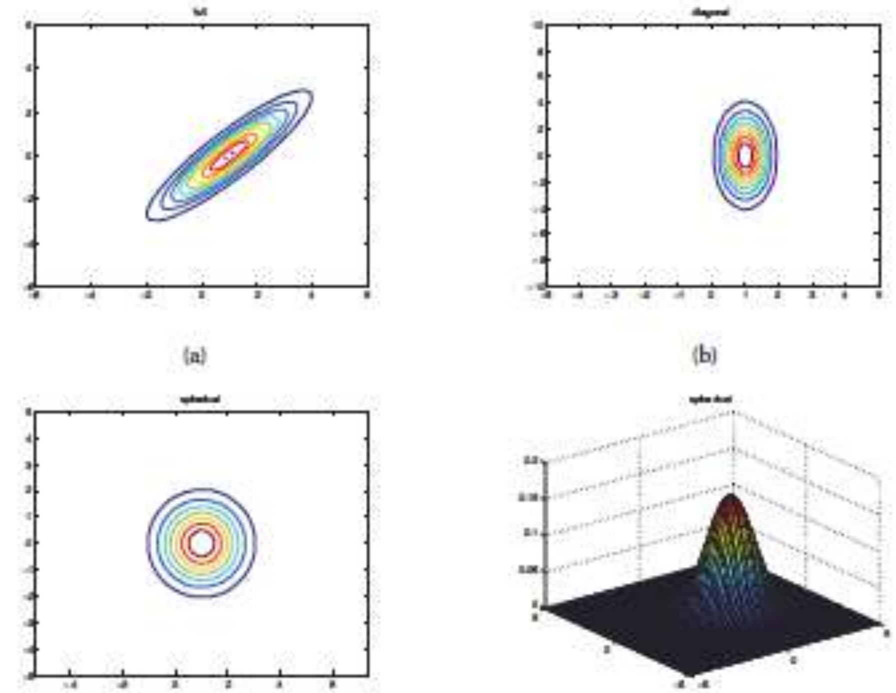


- $$\mathrm{cov}\,[\mathbf{x}] \triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}\,[\mathbf{x}])(\mathbf{x}-\mathbb{E}\,[\mathbf{x}])^T\right] = \begin{pmatrix} \mathrm{var}\,[X_1] & \mathrm{cov}\,[X_1,X_2] & \cdots & \mathrm{cov}\,[X_1,X_d] \\ \mathrm{cov}\,[X_2,X_1] & \mathrm{var}\,[X_2] & \cdots & \mathrm{cov}\,[X_2,X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}\,[X_d,X_1] & \mathrm{cov}\,[X_d,X_2] & \cdots & \mathrm{var}\,[X_d] \end{pmatrix}$$ $$\mathrm{corr}\,[X,Y] \triangleq \frac{\mathrm{cov}\,[X,Y]}{\sqrt{\mathrm{var}\,[X]\,\mathrm{var}\,[Y]}}$$
- corr [X, Y] = 1 iff Y = aX + b
- independence imply uncorrelation but uncorrelation does not imply independence
  - X = Unif[−1, 1] and Y = X^2

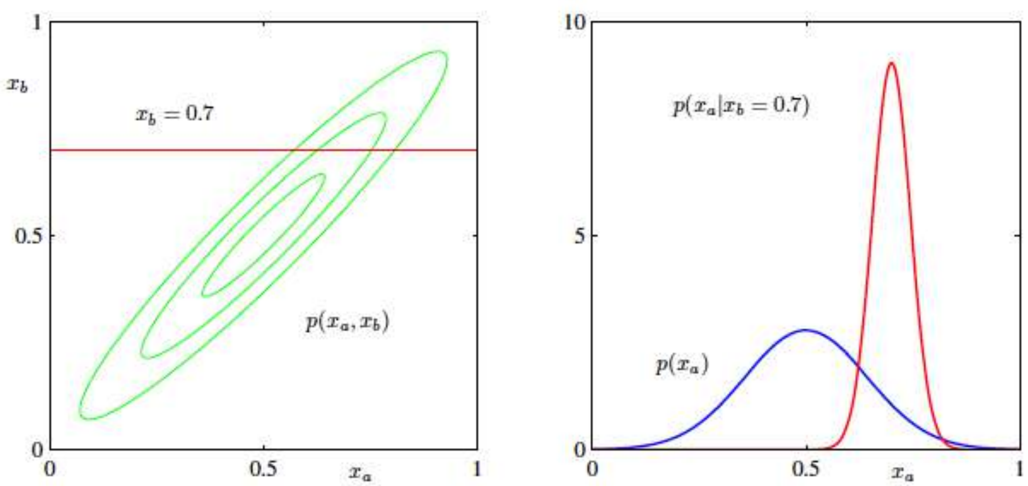


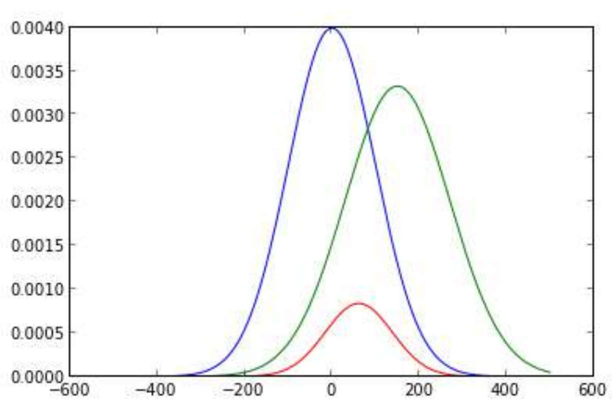
Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

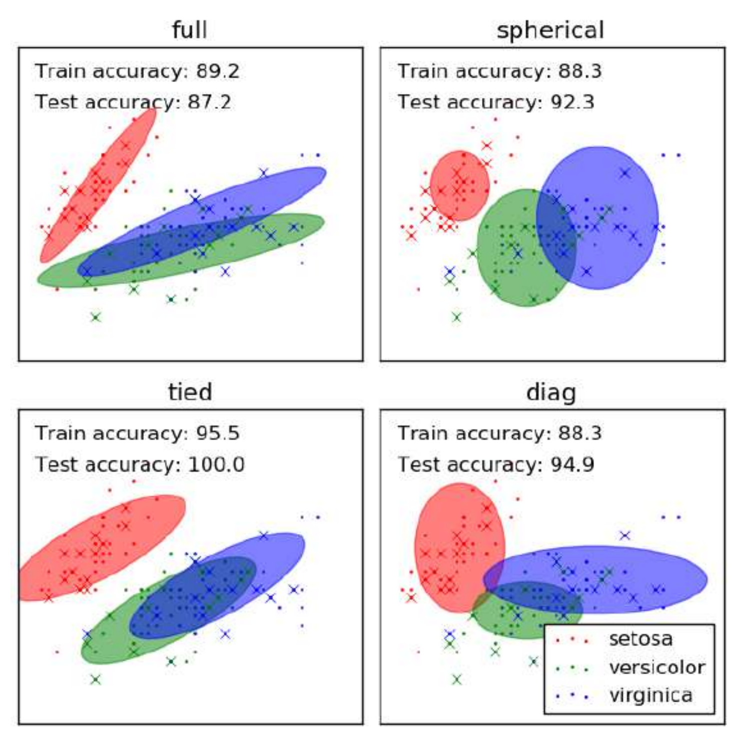


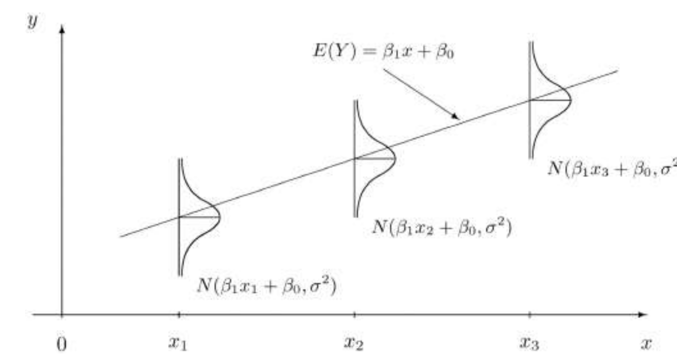
- Marginal and conditional distributions are Gaussian,
- product of Gaussians are Gaussian
- Gaussian mixture model

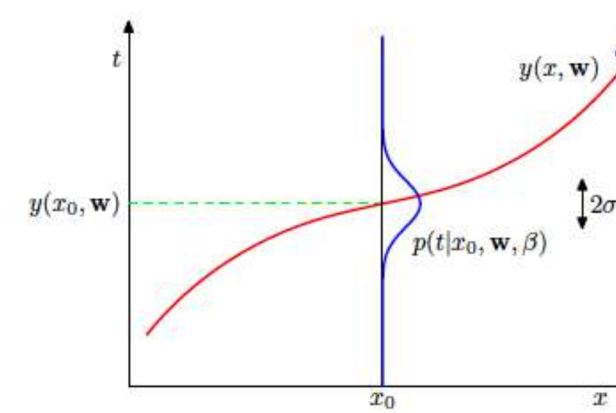




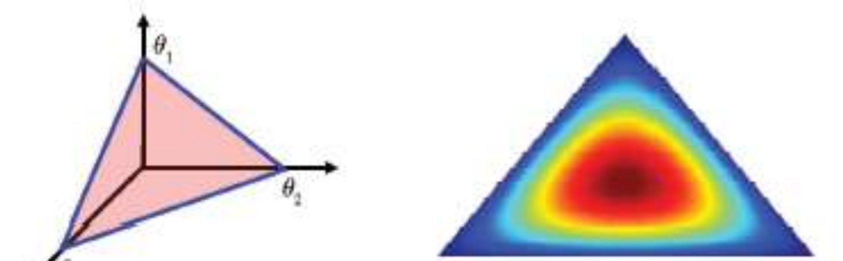
$$p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k)$$



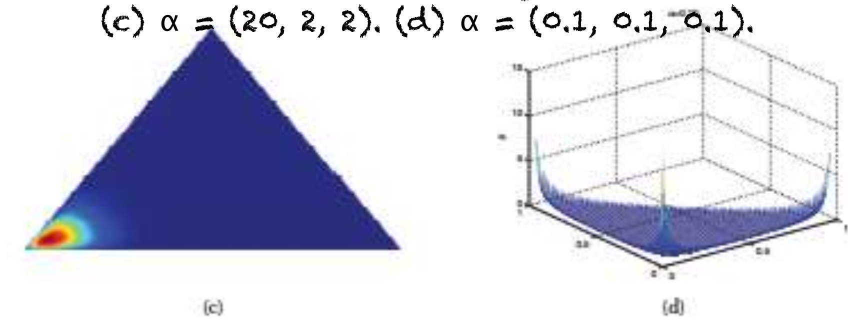




Dirichlet Distribution



(b) Plot of the Dirichlet density when α = (2, 2, 2).
(c) α = (20, 2, 2). (d) α = (0.1, 0.1, 0.1).



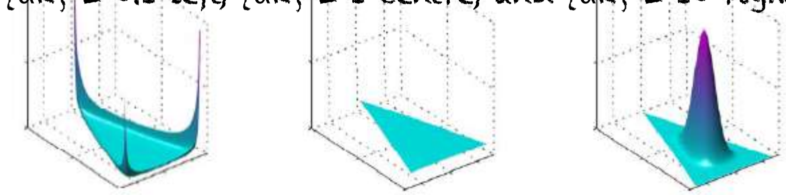
- Distribution over a probability simplex

$$S_K = \{\mathbf{x} : 0 \le x_k \le 1, \sum_{k=1}^{K} x_k = 1\}$$

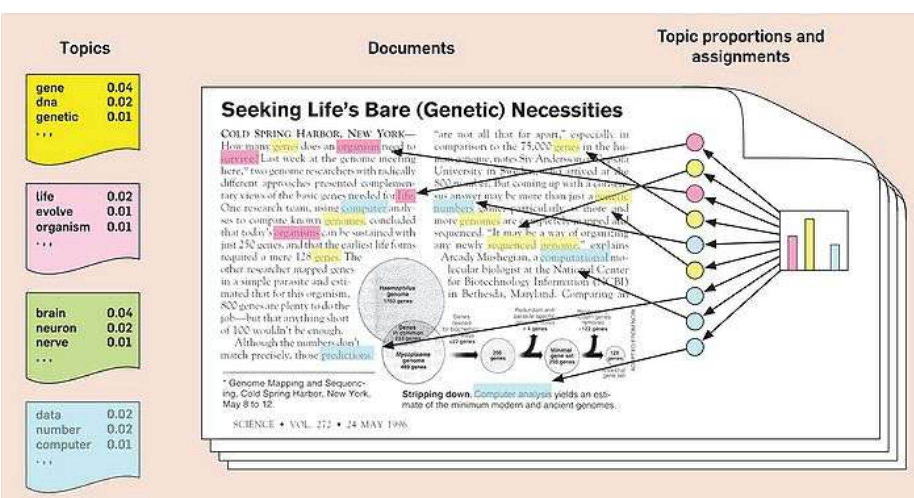
$$\mathrm{Dir}(\mathbf{x}|\boldsymbol{\alpha}) \triangleq \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K)$$

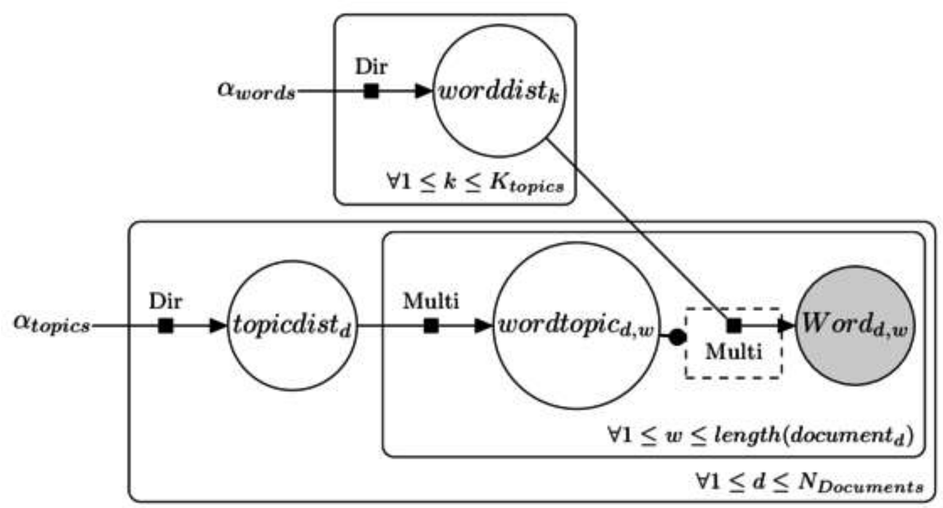
$$\mathbb{E}[x_k] = \frac{\alpha_k}{\alpha_0}, \ \mathrm{mode}[x_k] = \frac{\alpha_k - 1}{\alpha_0 - K}, \ \mathrm{var}[x_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

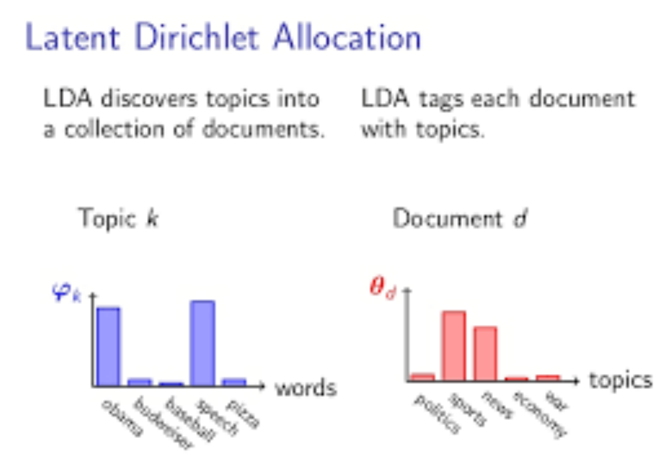
{αk} = 0.1 left, {αk} = 1 centre, and {αk} = 10 right



- Latent Dirichlet Allocation

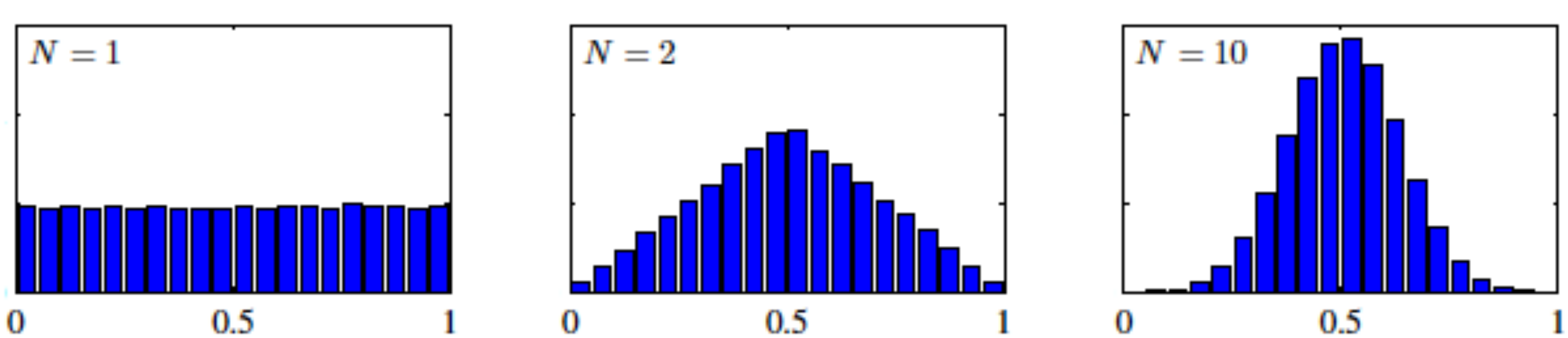


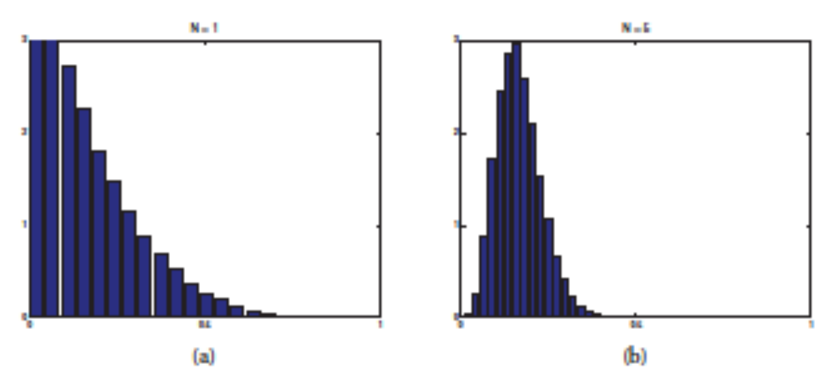




Central Limit Theorem

- Distribution of sum independent and identically distributed random variables $S_N = \sum_{i=1}^{N} X_i$ $\quad p(S_N = s) = \frac{1}{\sqrt{2\pi N \sigma^2}} \exp\left(-\frac{(s - N\mu)^2}{2N\sigma^2}\right)$
- Zn is standard normal $Z_N \triangleq \frac{S_N - N\mu}{\sigma\sqrt{N}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{N}}$ $\quad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- X is binomially distributed with parameters n and p, then X has the same distributionas the sum of n independent Bernoulli random variables, each with parameter p. $\frac{X - E[X]}{\sqrt{\mathrm{Var}(X)}} = \frac{X - np}{\sqrt{np(1-p)}}$
- X be the number of times that a fair coin, flipped 40 times, lands heads. Find the probability that X = 20.





Transformation of Random Variables

$$\mathbf{y} = f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} \qquad \mathbb{E}[\mathbf{y}] = \mathbb{E}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\text{cov}[\mathbf{y}] = \text{cov}[\mathbf{A}\mathbf{x} + \mathbf{b}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$$

- Y = f(X)

- X : Discrete ; px(X) is uniform on the set {1 , . . . , 10}, f(X) = 1 if X is even and f(X) = 0 otherwise

$$p_y(y) = \sum_{x:f(x)=y} p_x(x)$$

- X : Continous

$$P_y(y) \triangleq P(Y \le y) = P(f(X) \le y) = P(X \in \{x | f(x) \le y\})$$

- f : monotonic

$$P_y(y) = P(f(X) \le y) = P(X \le f^{-1}(y)) = P_x(f^{-1}(y))$$

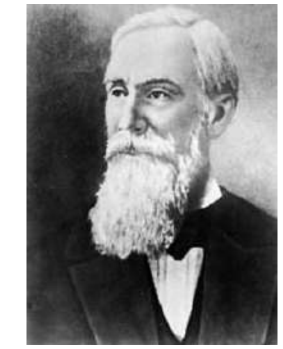
- $p_y(y) \triangleq \frac{d}{dy} P_y(y) = \frac{d}{dy} P_x(f^{-1}(y)) = \frac{dx}{dy}\frac{d}{dx} P_x(x) = \frac{dx}{dy} p_x(x)$ $\qquad p_y(y) = p_x(x) \left|\frac{dx}{dy}\right|$ $\qquad f_Y(y) = \sum_{x^2=y} f_X(x) \left|\frac{dx}{dy}\right|$

Transformation of Random Variables

- $X \sim U(-1, 1)$, and $Y = X^2$.

$$p_y(y) = p_x(x)\left|\frac{dx}{dy}\right|$$

Limit Theorems



- Markov Inequality : X is a random variable that takes only nonnegative values, then for any value a > 0

$$P\{X \geq a\} \leq \frac{E[X]}{a}$$

- X is a random variable with mean μ and variance σ2, then, for any k > 0

$$P\{|X - \mu| \geq k\} \leq \frac{\sigma^2}{k^2}$$

- Useful when only mean, or both the mean and the variance, and not distribution of X

Limit Theorems Example

- Suppose we know that the number of items produced in a factory during a week is a random variable with mean 500.
- (a) What can be said about the probability that this week's production will be at least 1000?
- (b) If the variance of a week's production is known to equal 100, then what can be said about the probability that this week's production will be between 400 and 600?

Entropy

- measure of its uncertainty

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k)$$

- [0. 25,0. 25,0. 2,0. 15,0. 15], [0. 2,0. 2,0. 2,0. 2,0. 2]
- maximum entropy is the uniform distribution
- compactly representing data(short codewords to highly probable bit strings)
- natural language, common words (“a”, “the”, “and”) are short
- Bernoulli r.v. for what value of θ, entropy is maximum ?
- Many models in ML such as MEMM, CRFs are based on maximum entropy principle - choose the simplest model

Kullback Leibler Divergence

- KL : measure the dissimilarity of two probability distributions
  - average number of extra bits needed to encode the data
- H(p, q) : cross entropy
- average number of bits to encode data with distribution p but using q

$$\mathbb{KL}(p||q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

$$\mathbb{H}(p,q) \triangleq -\sum_{k} p_k \log q_k$$

- (Information inequality) $KL(p||q) \geq 0$ with equality iff $p = q$.
- discrete distribution with the maximum entropy is the uniform distribution
- Learning and prediction in Bayesian models like LDA, Gaussian processes etc. and deep learning models such as variational auto encoders use KL

Mutual Information

- covariance captures only linear correlation
- Similar the joint distribution p(X, Y) is to the factored distribution p(X)p(Y)

$$\mathbb{I}(X;Y) \triangleq \mathbb{KL}(p(X,Y)||p(X)p(Y)) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \qquad \mathbb{H}(Y|X) = \sum_x p(x)\mathbb{H}(Y|X=x).$$

- reduction in uncertainty about X after observing Y
- Pointwise mutual information: discrepancy between events occuring together or by chance

$$\mathrm{PMI}(x,y) \triangleq \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}$$

- In NLP : if two words occur together or by chance

Parameter Estimation

- Maximum Likelihood estimation : argmax_θ p(x|θ) = argmax_θ log p(x|θ)
- Binary r.v.

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

- Multinomial

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

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


THANK YOU