Multiplication and permutation rules

sample k is drawn from a population of n distinc objects:

	order maters	order doesn't matter
replace	n^k	C_{n+k-1}^k
no replace	$n!$	C_n^k

Probability Axioms

$P(A \bigcup B \bigcup C) = P(A) + (B) + P(C) - P(A \bigcap B) - P(A \bigcap C) - P(B \bigcap C) + P(A \bigcap B \bigcap C)$

Conditional probability: $P(B|A) = P(A \bigcap B)/P(A)$
Independent events: two events are independent if:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \bigcap B) = P(A)P(B)$

Bayes theorem & Conditional probabilities

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

For hypothsis H and given data D, $P(H|D) = \dfrac{P(D|H)P(H)}{P(D)}$, P(H) is prior, which is unknown.
(Prior is important, see the cancer example with prevalence as 10^{-4} and 0.5.)

Simpson's paradox: one has higher success rate in every single operation but has overall success rate.
Monty Hall problem

Discrete Distribution

Random variable: X; measured value: x.
Probability mass function (**PMF**): $P(X = x)$

Cumulative distribution function (**CDF**): $P(X \leq x)$
Complementary cumulative distribution function (**CCDF**): $1 - CDF$.

mean: measure of center of mass; 1st moment; $\mu = E(X) = \sum_x x * P(X = x)$
variance: measure of dispersion; 2nd moment; $\sigma^2 = V(X) = \sum_x (x - \mu)^2 f(x) = E(x^2) - \mu^2$ (can be infinite: $P(X = x) \geq 1/x^3$)
skewness: how asymmetric is the distribution around the mean. Normalized 3-rd moment: $\gamma = E(\frac{(x-\mu)^3}{\sigma^3})$ (can be infinite: $P(X = x) \geq 1/x^4$)
geometirc mean: for very broad distribution. Mean is dominated by very unlikely but very large events (like lottery). It is $exp(E(logX))$.
NOTE: All can be infinite.

Discrete uniform distribution

$f(x) = 1/(b - a + 1)$, a, b is integer
$\mu = (b + a)/2$
$\sigma^2 = [(b - a + 1)^2 - 1]/12$

Bernouli distribution

$f(x) = p, if\ x = 1; 1 - p, if\ x = 0$
$E(X) = p; Var(X) = p(1 - p)$

Binomial distribution

sum of n independent bernouli trials, $f(x) = C_x^n p^x (1 - p)^{n-x}$

Poisson distribution

$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$
$E(X) = \lambda, Var(X) = \lambda$

- covered genome fraction: coverage=\lambda=NL/G; $P(X > 0) = 1 - exp(-\lambda)$
- how many configs: modified $\lambda = (N - 1)(L - L_{ov}/G)$, probability no left ends fall inside a read, $N_{config} = Nexp(-\lambda)$
- average length of config: $G_{covered}/N_{config}$

Geometric distribution

continue until success: $P(X=x) = p(1-p)^{x-1}$
$E(X) = 1/p; Var(X) = (1-p)/p^2$

- time to last common maternal ancestor: $P(T=t) = (1-1/N)^{t-1}(1/N)$

Negative binomial distribution

number of trials until r successes: $f(x) = C_{r-1}^{x-1} p^r (1-p)^{x-r}$
$E(X) = r/p; Var(X) = r(1-p)/p^2$

- cancer passenger and driver mutations

Power Law Distribution

$P(X=x) = Cx^{-\lambda}$, where C is normlization term, $1 = \sum_x C.x^{-\lambda}$ -> $C = 1/\zeta(\lambda)$. **Mean and variance** can be infinite.

- protein-protein network
- cancer mutation

Continuous Distribution

PDF is the derivative of CDF: $f(x) = \frac{dF(x)}{dx}$
$E(X) = \int_{-\infty}^{\infty} x f(x) dx; Var(X) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2$

Contunous uniform distribution

$f(x) = 1/(b-a)$
$E(X) = (b+a)/2; Var(X) = (b-a)^2/12$

Constant rate (poisson process)

Discrete events happen at rate r; expected #events in time x is rx.
The actual #events N_x is a poisson distribution discrete random variable. $p(N_x = n) = \frac{(rx)^n}{n!} exp(-rx)$

Divide x into many tiny intervals of length Δx, so $p(N=n) = C_n^l p^n (1-p)^{L-n}$, where $p \sim \Delta x = r\Delta x \to 0$ and $L \sim 1/\Delta x = x/\Delta x \to \infty$. Therefore, $E(N_x) = pL = rx$

Exponential Distribution

Models the time interval to the 1st event.

Exponential random variable X describes **interval** between 2 successes of a constant rate random process with success rate r per unit interval.

PDF: $f(x) = re^{-rx}, 0 \le x < \infty$

closely related to discrete geometric distribution: $f(x) = p(1-p)^{x-1} = pe^{(x-1)ln(1-p)} \approx pe^{-px}$ for small p.

X is continuous:

CCDF: $P_x(X > x) = P_N(N_x = 0) = exp(-rx)$

PDF: $f_x(x) = -dCCDF_X(x)/dx = r * exp(-rx)$

$u = E(X) = \frac{1}{r}$ and $\sigma^2 = V(X) = \frac{1}{r^2}$

Exponential distribution is the only memoryless distribution.

Proof: $P(x > t+s|x > s) = \frac{P(x > t+s, x > s)}{p(x > s)} = \frac{exp(-\lambda(t+s))}{exp(-\lambda s)} = exp(-\lambda t) = P(x > t)$

Erlang Distribution

generalization of exponential distribution.

Models the time interval to the k^{th} event, a sum of k exponentially distributed variables

$P(X > x) = \sum_{m=0}^{k-1} \frac{e^{-rx}(rx)^m}{m!} = 1 - F(x)$

$f(x) = F(x)' = \frac{r^k x^{k-1} e^{-rx}}{(k-1)!}$

Gamma Distribution

random variable x with PDF as $f(x) = \frac{r^k x^{k-1} e^{-rx}}{\Gamma(k)}$ has a gamma random distribution. If k is an positive integer, X has an Erlang distribution.

$\int_0^\infty f(x)dx = 1$ => $\Gamma(k) = \int_0^\infty r^k x^{k-1} e^{-rx} dx = \int_0^\infty y^{k-1} e^{-y} dy,\ where\ y = rx$

Properties of Gamma function:

- $\Gamma(1) = 1$
- $\Gamma(k) = (k-1)\Gamma(k-1)$, recursive property
- $\Gamma(k) = (k-1)!$, factorial function

- $\Gamma(1/2) = \pi^{1/2} = 1.77$

Mean and Variance of Erlang and Gamma:
$\mu = E(X) = k/r, \sigma^2 = V(x) = k/r^2$

Normal/Gaussian Distribution

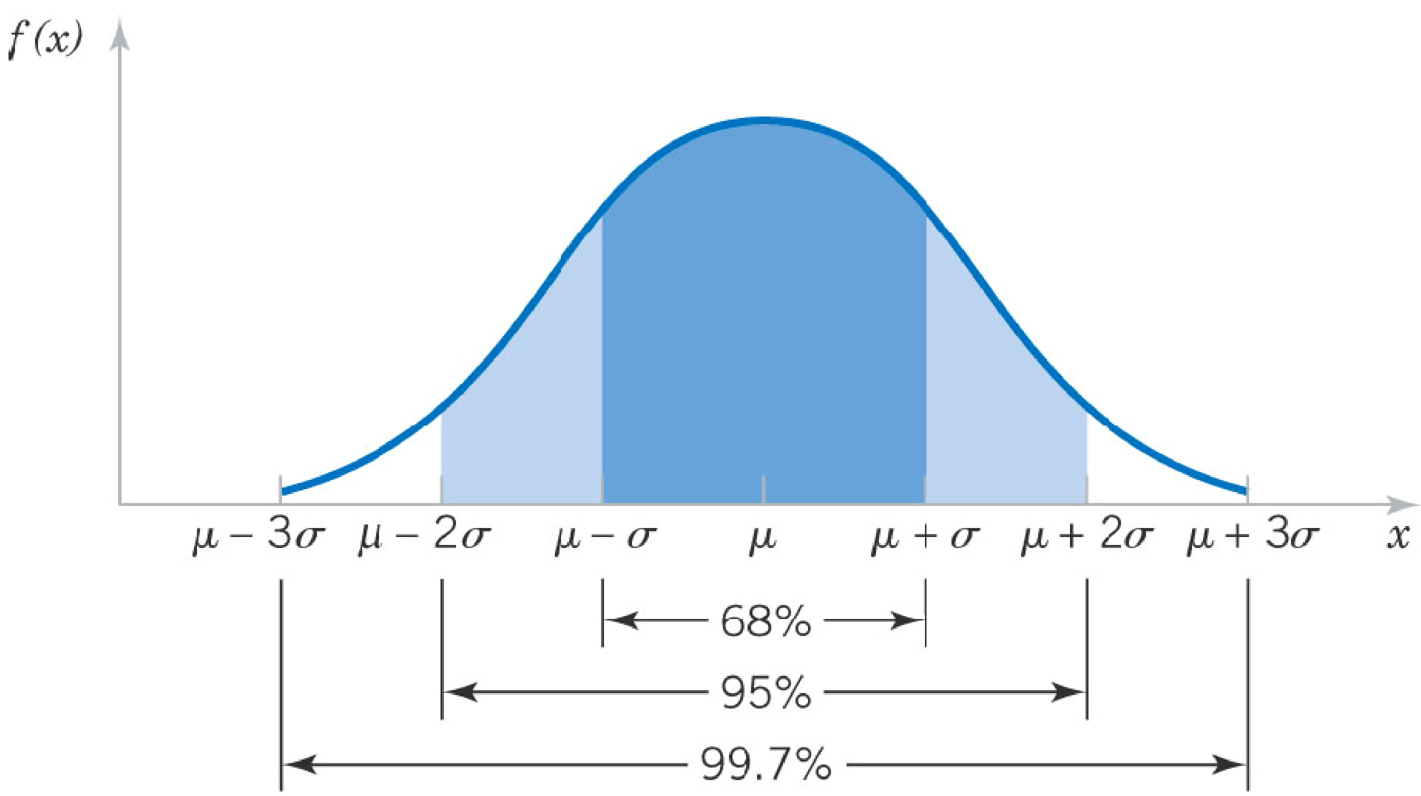
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \sim N(\mu, \sigma)$$

The sum of many independent random variables could be approximated with a Gaussian.

Standard Normal Distribution

$Z \sim N(0,1)$
CDF is $\Phi(z) = P(Z \leq z)$

$N \sim (\mu, \sigma)$ can be **standardized** into $N \sim (0,1)$ by $Z = \frac{X-\mu}{\sigma}$. => $P(X \leq x) = P(Z \leq z)$
$P(X < \mu - \sigma) = P(X > \mu + \sigma) = 0.16$
$(X < \mu - 2\sigma) = P(X > \mu + 2\sigma) = 0.023$
$(X < \mu - 3\sigma) = P(X > \mu + 3\sigma) = 0.0013$



CDF of normkal distribution in MATLAB:

1. erf function

```
(1-erf((x-u)/(sigma*sqrt(2)))/2
```

2. normcdf function

```
1-normcdf(x, 0, 1)
```

Lognormal Distribution

$X = e^W$, where $W\ N(\theta, \omega)$, => $W = ln(X)$
X is a lognoraml distribution variable.
$F(x) = P(X < x) = P(exp(W) \leq x) = P(W) \leq ln(x) = P(Z \leq \frac{ln(x) - \theta}{\omega}) = \Phi(\frac{ln(x) - \theta}{\omega})$ for $x > 0$; or 0 if $x \leq 0$
$f(x) = \frac{dF(x)}{dx} = \frac{1}{x}\frac{1}{\omega\sqrt{2\pi}}exp(-(\frac{ln(x) - \theta}{2\omega})^2)$ for $x > 0$
$E(X) = e^{\theta+\omega^2/2}$ and $V(X) = e^{2\theta+\omega^2}(e^{\omega^2} - 1)$

Joint Probability Distribution

Joint PMF, $f_{XY}(x, y)$
Marginal probability distribution

- $f_X(x) = \sum_y f_{XY}(x, y)$
- $f_Y(y) = \sum_x f_{XY}(x, y)$
  Use marginal distributions to compute E and V:

y = number of times city name is stated	x = number of bars of signal strength					
	1	2	3	$f(y)$ =	$y*f(y)$ =	$y^2*f(y)$ =
1	0.01	0.02	0.25	0.28	0.28	0.28
2	0.02	0.03	0.20	0.25	0.50	1.00
3	0.02	0.10	0.05	0.17	0.51	1.53
4	0.15	0.10	0.05	0.30	1.20	4.80
$f(x)$ =	0.20	0.25	0.55	1.00	2.49	7.61
$x*f(x)$ =	0.20	0.50	1.65	2.35		
$x^2*f(x)$ =	0.20	1.00	4.95	6.15		

$E(X) = 2.35; V(X) = 6.15 - 2.35^2$
$E(Y) = 2.49; V(X) = 7.61 - 2.49^2$

Conditional probability distribution
$P(Y = y|X = x) = P(X = x, Y = y)/P(X = x) = f(x, y)/f_X(x)$

Random variables independent if **all events** A that Y=y and B that X=x are independent if **any one** of the conditions is met:

- $P(Y = y|X = x) = P(Y = y)$
- $P(X = x|Y = y) = P(X = x)$
- $P(X = x, Y = y) = P(X = x).P(Y = y)$ for every pair of x and y

Conditional probability density function: $f_{Y|x}(y) = \dfrac{f_{XY}(x, y)}{f_X(x)}$

Independence of continuous random variable:

- $f_{XY}(x, y) = f_X(x)f_Y(y)$
- $f_{Y|x}(y) = f_Y(y); f_{X|y} = f_X(x)$
- $P(X \subset A, y \subset B) = P(X \subset A)P(Y \subset B)$

Covariance & Correlation

Covariance: measure dependence between random varibales
$Cov(X, Y) = \delta_{XY} = E(X, Y) - \mu_X \mu_Y \in (-\infty, \infty)$
If independent, $Cov(X, Y) = 0$. $\rho_{XY} = 0$ is necessary for independence, but not sufficient.

Correlation:

Pearson correlation: normalized covariance to test linear relationship between X and Y, unlikely for broad distribution.
$\rho_{XY} = \sigma_{XY}/\sigma_X \sigma_Y \in [-1, 1]$

Spearman rank correlation: test monotonic relationship between X and Y.
Calculate ranks (1 to n), $r_X(i)$ and $r_Y(i)$, $Spearman(X, Y) = Pearson(r_X, r_Y)$

Linear functions of random variables

$Y = c_1 X_1 + c_2 X_2 + ... + C_p X_p$
$E(Y) = c_1 E(X_1) + ... + c_p E(X_p)$
$V(Y) = c_1^2 V(X_1) + c_p^2 V(X_p) + 2 \sum_{i<j} \sum c_i c_j cov(X_i X_j) = c_1^2 V(X_1) + ... + c_p^2 V(X_p)$, if $cov(x_i, x_j) = 0$

Average $\overline{X} = (X_1 + X_2 + ..X_p)/p$, then $E(\overline{X}) = \mu; V(\overline{X}) = \delta^2/p$

PCA

diagonalize the $p \times p$ matrix of correlation coefficients $r_i j = \dfrac{\sigma_{ij}}{\sigma_i \sigma_j}$. Produce new variable $y_1, y_2, ..., y_p$ from original $x_1, x_2, ..., x_p$.
$y_i = a_{i1} x_1 + a_{i2} x_2 + ... + a_{ip} x_p$