	Name	Description	Support	PMF	CDF	E[X]	$E[X^2]$	var(X)	Transform
DISCRETE	Bernoulli	1 success, 0 failure	{0,1}	$P_X(1)=p$, $P_X(0)=1-p$	0, steps to (1-p) at 0, steps to (1) at 1.	p	p	p(1-p)	$1 - p + pe^s$
	Binomial	# successes in n Bernoullis*	0<k<n	$P_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$	$(n-k)\binom{n}{k}\int_0^{1-p} t^{n-k-1}(1-t)^k dt$	np	np(np-p+1)	np(1-p)	$(1 - p + pe^s)^n$
	Geometric	# trials to get first success (inclusive)	k ϵ I > 0	$P_X(k) = (1-p)^{k-1}p$	$1-(1-p)^k$	1/p	$(2-p)/p^2$	$(1-p)/p^2$	$\frac{pe^s}{1-(1-p)e^s}$
	Poisson	# rare events; approximates binomial with λ=np when n da, p xiao (cont-time version of binomial, though the answer is still discrete)	k ϵ I >= 0	$P_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for k = 0,1,2….	ugly	λ	$\lambda + \lambda^2$	λ	$e^{\lambda(e^s-1)}$
	Poission(k,t)	prob there are exactly k arrivals in t time		$P_X(k,t) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$	ugly	λt	$\lambda t + \lambda^2 t^2$	λt	
	Uniform	an integer in the interval [a,b]	k ϵ I, a<= k<=b	$P_X(k) = 1/(b-a+1)$	(floor(k) – a + 1)/n, for k between a and b inclusive. n = b-a+1	(a+b)/2	$(4a^2+3b^2+4ab+2b-2a)/12$	$\frac{(b-a)(b-a+2)}{12}$	$\frac{e^{sa}(e^{s(b-a+1)}-1)}{(b-a+1)(e^s-1)}$
	Pascal	time to kth arrival in a Bernoulli process	t>=k, where k is fixed	$P_{Y_k}(t) = \binom{t-1}{k-1}p^k(1-p)^{t-k}$	ugly	$E[Y_k] = k/p$	$k(k+1-p)/p^2$	$var(Y_k) = k(1-p)/p^2$	
	Name	**Description**	**Support**	**PDF**	**CDF**	**E[X]**	$\mathbf{E[X^2]}$	**var(X)**	**Transform**
CONTINUOUS	Uniform	a real number in [a,b]	a<=x<=b	1/(b-a)	$\frac{x-a}{b-a}$ btw a, b x<a,0 x>b,1	(a+b)/2	$(a^2+b^2+ab)/3$	$(b-a)^2/12$	$\frac{e^{sb}-e^{sa}}{s(b-a)}$
	Exponen'l	Time to first success (cont version of geometric)	x>=0	$\lambda e^{-\lambda x}$	$1-e^{-\lambda x}$, x>=0	1/ λ	$2/\lambda^2$	$1/\lambda^2$	$\frac{\lambda}{\lambda-s}$ for s<λ
	Normal	Gaussian w/ mean μ, var σ^2	x ϵ R	$\frac{1}{\sqrt{2\pi}*\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	Use lookup table	μ	$\sigma^2+\mu^2$	σ^2	$e^{\left(\frac{\sigma^2 s^2}{2}\right)+\mu s}$
	Soarnorv	$Y = X_1 + \ldots + X_N$, Xs iid and N indep from all of them				E[Y]=E[N] E[X]	var + $(E[X])^2$, it doesn't simplify	var(Y) = E[N]var(X) + $(E[X])^2$var(N)	$M_Y(s) = M_N(\ln M_X(s))$ aka find M_N and replace each e^s with $M_X(s)$
	Erlang 二郎	Time of kth arrival (cont version of pascal)	y > 0	$f_{Y_k}(y) = \frac{\lambda^k y^{k-1}e^{-\lambda y}}{(k-1)!}$	ugly	$E[Y_k] = kE[T] = k/\lambda$	$(k^2+k)/\lambda^2$	$var(Y_k) = kvar(T) = k/\lambda^2$	

Think about rearranging these things:

Law of conditional variances: var(X) = E[var(X|Y)] + var(E[X|Y])
- note $var_X(X|Y)$ is an RV, func of Y

Law of iterated expectation: E[X] = E[E[X|Y]]
- note $E_X[X|Y]$ is an RV, a func of Y

Variance: var(X) = $E[X^2] - (E[X])^2$

Total expectation theorem: E[X] = E[X|A]P(A) + E[X|B]P(B), where A, B are disjoint events covering all of sample space

LMS

Minimizes $E[(Y-g(X))^2]$ to var(Y|X)

$\hat{Y} = g(X) = E[Y|X]$

(BTW: LMS w/ no information, i.e. no conditioning, is just E[X] and error is var(X))

LLMS

Minimizes $E[(Y-aX-B)^2]$ to $(1-\rho^2)\sigma_Y^2$

$\hat{Y} = l(X) = aX + B$

$\hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)}(X - E[X])$

COUNTING	Ordered ("Tuplets")	Unordered ("Sets")
w/ replacement	n^k	$\binom{n+k-1}{k}$
w/o replacement	$\frac{n!}{(n-k)!}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Covariance
cov(X,Y) = E[(X-E[X])(Y-E[Y])] = E[XY] – E[X]E[Y].
if cov(X,Y)==0, then X and Y are uncorrelated.
cov positive means “tend” to have same sign, i.e. xy plot is increasing, negative means “tend” to have opposite sign,i.e. xy plot is decreasing
cov(X,X) = var(X)
cov(X,aY+b)=a*cov(X,Y)
cov(X,Y+Z)=cov(X,Y)+cov(X,Z)

Correlation Coefficient ρ
Normalized version of cov, ϵ [-1,1]
ρ(X,Y) = $\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}}$

Var(X_1+X_2), X_1,X_2 not nec. indep
= var(X_1) + var(X_2) + 2cov(X_1,X_2)

$(af+bg)' = af' + bg'$
$(fg)' = f'g + fg'$
$\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$
$\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}.$
$\frac{d}{dx}\ln(x) = \frac{1}{x}, \qquad x > 0$
$\frac{d}{dx}a^x = \ln(a)a^x$
$\frac{d}{dx}e^x = e^x$

Stuff about Gaussians
N(0,1): $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
X~N(μ, σ^2), Y=aX+b 则 Y~N(aμ+b ,$a^2\sigma^2$)
X~N(μ, σ^2), P(X<=x) = Φ((x-μ)/σ)
If X, Y Gaussian and indep, W=X+Y is also Gaussian

Definitions
E[X] = $\int_{-\infty}^{\infty} x f_X(x)dx$
E[g(X)] = $\int_{-\infty}^{\infty} g(x)f_X(x)dx$
var(X) = $\int_{-\infty}^{\infty}(x-E[X])^2 f_X(x)dx$
= E[X^2] – (E[X])^2
E[aX+b]=aE[X]+b
var(aX+b)=a^2var(X)

Definitions
F_X(x) = P(X<=x)
continuous: = $\int_{-\infty}^{x} f_X(t)dt$
f_X(x) = $\frac{dF_X}{dx}(x)$
discrete: = $\sum_{k\le x} p_X(k)$
p_X(k) = F_X(k)-F_X(k-1)

$\sum_{k=0}^{n} ar^k = \frac{a(r^{n+1}-1)}{r-1}$
$\sum_{k=m}^{n} ar^k = \frac{a(r^{n+1}-r^m)}{r-1}$

Conditioning
$f_{X|Y}$(x|y) = $f_{X,Y}$(x,y)/f_Y(y)
f_X(x) = $\int_{-\infty}^{\infty} f_Y(y)f_{X|Y}(x|y)dy$
P(A) = $\int_{-\infty}^{\infty} P(A|X=x)f_X(x)dx$

Derived distributions
Y=g(X) get CDF: F_Y(y) = P(Y<=y)=P(g(X)<=y)=$\int_{x|g(x)\le y} f_X(x)dx$
then differentiate: $\frac{dF_Y}{dy}$(y)= f_Y(y)
also if Y=aX+b then f_Y(y) = $\frac{1}{|a|}f_X(\frac{y-b}{a})$

Continous Bayes’ Rule
X,Y cont, N disc, A an event
.$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_{-\infty}^{\infty} f_{Y|X}(y|t)f_X(t)dt}$
.$P(A|Y=y) = \frac{P(A)f_{Y|A}(y)}{f_Y(y)} = \frac{P(A)f_{Y|A}(y)}{f_{Y|A}(y)P(A)+f_{Y|Ac}(y)P(Ac)}$
.$P(N=n|Y=y) = \frac{p_N(n)f_{Y|N}(y|n)}{f_Y(y)} = \frac{p_N(n)f_{Y|N}(y|n)}{\sum_i p_N(i)f_{Y|N}(y|i)}$

$\int e^u du = e^u + C$
$\int a^u du = \frac{a^u}{\ln a} + C$
$\int ue^u du = e^u(u-1) + C$
$\int u^n e^u du = u^n e^u - n\int u^{n-1}e^u du$
$\int u^n a^u du = \frac{u^n a^u}{\ln a} - \frac{n}{\ln a}\int u^{n-1}a^u du + C$
$\int \frac{e^u}{u^n}du = -\frac{e^u}{(n-1)u^{n-1}} + \frac{1}{n-1}\int \frac{e^u}{u^{n-1}}du$
$\int \frac{a^u}{u^n}du = -\frac{a^u}{(n-1)u^{n-1}} + \frac{\ln a}{n-1}\int a^u u^{n-1}du$
$\int \ln u\, du = u\ln u - u + C$

Integration by parts:
$$\int u\, dv = uv - \int v\, du.$$

Convolution
W=X+Y, X,Y indep
Discrete: p_W(w)=$\sum_x p_X(x)p_Y(w-x)$
Continuous: f_w(w)=$\int_{-\infty}^{\infty} f_X(x)f_Y(w-x)dx$
Or use transforms: M_W(s)=M_X(s)M_Y(s)

Marginal pdf from a joint pdf
f_X(x) = integral[$f_{X,Y}$(x,y)dy]
you integrate with respect to y because you want to get rid of y and just have x

Geometric series
1/(1-a) = 1 + a + a^2 + …. out to infinity, for |a|<1

Transforms
M_X(s) = E[e^{sX}]
example of a transform:
p_X(x) = { 1/2, x=2; 1/6, x=3; 1/3, x=5}. M_X(s) = (1/2)exp(2s) + (1/6)exp(3s) + (1/3)exp(5s)
because of the definition of expectation. you know, M_X(s) =

How to do integrals on TI-83
MATH->fnInt(function,X,lowerlimit,upperlimit)
ex. fnInt(X^2,X,0,2) = 2.666667
similarly, nDeriv(X^2,X,3) = 6

summation(e^{sx}*p_X(x) for all x)
IF X and Y are independent, (otherwise the second-to-last step doesn't hold)
Z=X+Y $M_Z(s) = E[e^{sZ}] = E[e^{s(X+Y)}] = E[e^{sX}e^{sY}] = E[e^{sX}]E[e^{sY}] = M_X(s)M_Y(s)$
addition of indep RVs corresponds to multiplication of transforms

Bernoulli processes
$Y_k=T_1+\ldots+T_k$, Y_k is time of kth arrival, T_is are times between arrivals
Each time slot: Bernoulli.
Time til next arrival: Gemoetric.
#arrivals in interval: Binomial.
Time of kth arrival: Pascal.

Bernoulli processes contd
Splitting a p process by means of a q process: a pq process and a p(1-q) process
Merging a p process and a q process: a p+q-pq process. (-pq removes the double counting)

Balance equations for Markov Chain analysis
π: steady state probabilities. defined as long as no reachable recurrent states are periodic.
$\pi_j = \Sigma_i \pi_i p_{ij}$ (sum over *incoming* arrows)
$1 = \Sigma_i \pi_i$
(implies $1 * \pi^T = \pi^T * p$)
μ: time to absorption in class A. defined when finite
$\mu_i = 0$, i ϵ A
$\mu_i = 1 + \Sigma_j p_{ij} \mu_j$ o.w. (sum over *outgoing* arrows)
a: absorption probability into class A
$a_i = 0$, i ϵ C where C is a separate recurrent class
$a_i = 1$, i ϵ A
$a_i = \Sigma_j p_{ij} a_j$ o.w. (sum over *outgoing* arrows)

Comparison of Poisson and Bernoulli processes

	Poisson	Bernoulli
Time of arrival	Continuous	Discrete
Inter-arrival time	Exponential	Geometric
Number of arrivals	Poisson	Binomial
Time of kth arrival	Erlang	Pascal

Poisson processes
When you merge them you get a new Poisson process with parameter = sum of the original parameters. Time to first success is exponential.

Merge/split Poisson processes
Merge: new poisson, parameter is sum
Split w/ prob q and (1-q) results in two poisson processes of rate λq and λ(1-q)

Markov Chains
probability of state j next is same whether conditioned on current state or on entire history so far; in other words, memory is just one step deep

Frequency interpretation of steady state probs
Long-run frequency of being in state j is π_j
Frequency of transitions j->k is $\pi_j p_{jk}$ (useful for birth-death processes)

Absorption
A recurrent state k is absorbing if probability of looping back to self is 1 and probability of going to any other state is 0.

Markov Chain Transition Matrices
Transition matrices are written
p11 p12
p21 p22
i.e. each row is one i, and each column is one j for pij.
rows must sum to 1; columns needn't

Limit Theorems
If you see "deviation far from mean," think Chebyshev
If asked to prove a sample average converges in probability, think WLLN

Markov Inequality
P(X>=a)<=E[X]/a, X always >=0
Chebyshev Inequality
$P(|X-E[X]|>=c)<=var(X)/c^2$
or,
$P(|X-E[X]|>=k\sigma)<=1/k^2$

Convergence in Probability
For every ϵ > 0,
lim n->inf P(|Yn-a|> ϵ) = 0
(note ϵ *strictly* > 0, NOT >=0)

Weak Law of Large Numbers
For every ϵ > 0, X_i iid, finite mean and variance, and define $M_n = (X_1+\ldots+X_n)/n$
lim n->inf $P(|M_n-\mu|>= \epsilon) = 0$ i.e. converges in prob to μ
This is a special case of Chebyshev, b/c that would give on right side $\sigma^2/n\epsilon^2$ which -> 0 as n->inf

Pollster's Problem
f = true fraction of population that support such-and-such
want $P(|M_n-f|>=.01)<=.05$
Chebyshev: $P(|M_n-\mu|>= \epsilon) <= \sigma^2/n\epsilon^2$
plug in ϵ = .01, worst case scenario $\sigma^2 = 1/4$ (max variance for a Bernoulli), then to get right side of equation down to .05 you need n >= 50,000

Central Limit Theorem
X_i iid with finite mean μ and var σ^2.
Define $Z_n = \frac{\sum_{i=1}^{n} X_n - n\mu}{\sigma\sqrt{n}}$ to normalize it
As n->inf, Z_n converges to CDF of ~N[0,1] a.k.a. Φ(z)
So lim n-> inf $P(Z_n<=z) = \Phi(z)$

Remember you can subtract CDFs to get the probability of being between two numbers

Central Limit Theorem
central limit theorem (CLT) states conditions under which the sum of a sufficiently large number of independent random variables, each with finite mean and variance, will be approximately normally distributed
$P(S_n<=c)$ ~= Φ(z), where $z = \frac{c-n\mu}{\sigma\sqrt{n}}$, and S_n has mean nμ and var $n\sigma^2$
i.e. μ and σ refer to the mean and stddev of the ORIGINAL vars, not their sum

Maximum Likelihood Estimation: Intro
Distribution $p_x(x;\theta)$ has unknown parameter set θ which is not random. We know nothing about what it might be (this is classical statistical inference)
Want a point-estimate of θ
For each value θ you would get a different PDF or PMF and thus a different likelihood of the observed vector $X = (X_1,\ldots,X_n)$
If the X_i are iid, then the joint PDF of all them is the product of the individual PDFs. So its log is the sum of their logs.

Maximum Likelihood Estimation: More
You want to pick the value of θ that gives the largest likelihood of what you observed, i.e.
thetahat = $\text{argmax}_\theta p_x(x;\theta)$
Standard trick that works if something (?) is monotonically increasing:
thetahat = $\text{argmax}_\theta \ln p_x(x;\theta)$
set 0 = [d/dθ] of ln $p_x(x;\theta)$

Properties that an estimator can have
Parameter (vector) to be estimated: θ
Observation (vector) to use: X
Estimator: g(X)
Unbiased: E[g(X)] = θ
Asymptotically unbiased: lim n->inf $E[g(X_1,\ldots,X_n)]=\theta$
Consistent: $g(X_1,\ldots,X_n)$ converges to θ in probability as n -> inf

Binary Hypothesis Testing

Two hypotheses H_0 and H_1, one of which is true.
H_0 : null hypothesis. $p_x(x|H_0)$
H_1 : alternative hypothesis. $p_x(x|H_1)$
We observe a value X = x, giving us "likelihoods" of H_0 and H_1

Likelihood Ratio

The optimal way to do binary hypothesis testing.
$L(x) = p_X(x|H_1)/p_X(x|H_0)$
Likelihood ratio test (LRT):
L(x) ≷ 开 where > is H_1 and < is H_0
Target α = P(false rejection of null hypoth)
Choose 开 such that P(L(x)>开|H_0) = α
Reject H_0 if L(x) > 开

Neyman-Pearson Lemma

For given α (false rejection), LRT gives smallest possible β (false acceptance)

Types of error

R is rejection [of H_0] region
Type I : $\alpha = P(X \in R|H_0)$
Type II: $\beta = P(X \notin R|H_1)$ (that's a crossed-out epsilon)

Overall probability of an error

$P(\text{error}) = P(H_0)\alpha + P(H_1)\beta$, where $P(H_0)=P_0$ and $P(H_1)=P_1$ are a priori probabilities.
P(error) = P_0 integral$_R$($f_X(x|H_0)$dx) + P_1 *[1- integral$_R$($f_X(x|H_1)$dx)]
P(error) = P_1 + integral$_R$ [$P_0 f_X(x|H_0) - P_1 f_X(x|H_1)$]dx
Sometimes you want to choose R to minimize P(error)

Overall error cont'd

L(x) = $f(x|H_1)/f(x|H_0)$ ≷ P_0/P_1 = 开
Where top (>) goes to H_1 and bottom (<) goes to H_0.
Or sometimes you minimize a cost function
cost = $c_1\alpha + c_2\beta$

Example of convolution from spring 2008 Q2

c. Suppose X is uniformly distributed over $[0, 4]$ and Y is uniformly distributed over $[0, 1]$. Assume X and Y are independent. Let $Z = X + Y$. Then

(i) $f_Z(4.5) = 0$
(ii) $\boxed{f_Z(4.5) = 1/8}$
(iii) $f_Z(4.5) = 1/4$
(iv) $f_Z(4.5) = 1/2$

Solution: Since X and Y are independent, the result follows by convolution:

$$\boxed{f_Z(4.5) = \int_{-\infty}^{\infty} f_X(\alpha) f_Y(4.5-\alpha)\, d\alpha = \int_{3.5}^{4} \frac{1}{4}\, d\alpha = \frac{1}{8}\,.}$$

Confidence Intervals

θ^- = g(X_1,…,X_n)
θ^+ = h(X_1,…,X_n)
You want P(θ^- <= θ <= θ^+) >= .95
suppose X_i Gaussian, known var v and unknown mean θ
θ^$_n$ = summation$_i$ (X_i)/n
Z_n = [θ^$_n$ – θ]/sqrt(v/n) 基本上 ~N[0,1]
P(|Z_n|<1.96) >= .95 implies P(θ^- <= θ <= θ^+) >= .95
θ^- = θ^$_n$ – 1.96sqrt(v/n)
θ^+ = θ^$_n$ + 1.96sqrt(v/n)
all because Φ(1.96) = .9750 so with both tails you get .95 between.

Misc tricks and giggles

For X_i iid, unknown mean μ and variance σ^2
$M_n = (X_1+\ldots+X_n)/n$ is an unbiased & consistent estimator of μ
V_n = summation$_i$ [$(X_i – M_n)^2$]/(n-1) is unbiased
You might be tempted to use
V_n = summation$_i$ [$(X_i – M_n)^2$]/n which is the MLE but is biased but is asymptóticamente unbiased.

Shit I don't think actually matters: t-distributions

Again, for unknown mean and variance we can define
$T_n = (M_n – \theta)$/sqrt(V_n/n)
T_n is not quite a true normal distribution, because M_n or V_n could differ from true μ and v.
Instead it is a t-distribution with n-1 degrees of freedom.
Incidentally, somehow, this distribution does not depend on θ or v. For n large (like > 50 say), the t-dist becomes very close to the standard normal dist.

Chernoff-Hoeffding Bounds, which totally suck

X_i iid, bounded in [a,b]. For all ϵ > 0,
P(|M_n – μ|>= ϵ) <= $2e^{-n\epsilon^2/(a-b)^2}$
Special case: bernoulli, so a and b are 0 and 1
P(|M_n – μ|>= ϵ) <= $2e^{-n\epsilon^2}$
Advantages: like the CLT and unlike Chebyshev, probability falls off exponentially. But unlike the CLT and like Chebyshev, it gives a bound rather than an approximation—"less handwaving".

Chernoff-Hoeffding Bounds cont'd

So, to use it, say you have n = 1000 and you need 95% confidence, then solve $2e^{-n\epsilon^2}$=.05 to find ϵ.
Then add and subtract ϵ from M_n to get an upper and lower bound with 95% confidence.
Advantage: uses no approximations; the bound is always correct.
Disadvantage: a looser confidence interval than some other things.

Theorem of Total Probability

$p_T(t) = \int_{-\infty}^{\infty} p_{T|Q}(t|q) f_Q(q)\,dq$
for instance,
$p_T(t) = \int_0^1 (1-q)^{t-1} q * 1\,dq$
where 0 to 1 is q's range, and the stuff left of * is conditional dist of t, and 1 is prob density of q.

Check your Markov Chain!

Is it valid? For each ball, do all outgoing arrows sum to 1?