Bernoulli trials

An experiment, or trial, whose outcome can be classified as either a success or failure is performed.

$X =$ 1 when the outcome is a success
0 when outcome is a failure

If p is the probability of a success then the pmf is,
$p(0) = P(X=0) = 1-p$ $p(1) = P(X=1) = p$

A random variable is called a Bernoulli random variable if it has the above pmf for p between 0 and 1.

Expected value of Bernoulli r. v.:

$E(X) = 0*(1-p) + 1*p = p$

Variance of Bernoulli r. v.:

$E(X^2) = 0^2*(1-p) + 1^2*p = p$
$\text{Var}(X) = E(X^2) - (E(X))^2 = p - p^2 = p(1-p)$

Ex. Flip a fair coin. Let X = number of heads. Then X is a Bernoulli random variable with $p=1/2$.
$E(X) = 1/2$
$\text{Var}(X) = 1/4$

Binomial random variables

Consider that n independent Bernoulli trials are performed. Each of these trials has probability p of success and probability (1-p) of failure.

Let X = number of successes in the n trials.

$p(0) = P(0$ successes in n trials$) = (1-p)^n$	{FFFFFFF}
$p(1) = P(1$ success in n trials$) = (n\ 1)p(1-p)^{n-1}$	{FSFFFFF}
$p(2) = P(2$ successes in n trials$) = (n\ 2)p^2(1-p)^{n-2}$	{FSFSFFF}

……
$p(k) = P(k$ successes in n trials$) = (n\ k)p^k(1-p)^{n-k}$

A random variable is called a Binomial(n,p) random variable if it has the pmf,
$p(k) = P(k$ successes in n trials$) = (n\ k)p^k(1-p)^{n-k}$ for k=0,1,2,….n.

Valid pmf: sum_{k=0}^n p(k) = 1
sum_{k=0}^n (n k)$p^k(1-p)^{n-k} = (p+(1-p))^n = 1$

Ex. A Bernoulli(p) random variable is binomial(1,p)

Ex. Roll a dice 3 times. Find the pmf of the number of times we roll a 5.

X = number of times we roll a 5 (number of successes)
X is binomial(3,1/6)

$p(0) = (3\ 0)(1/6)^0(5/6)^3 = 125/216$
$p(1) = (3\ 1)(1/6)^1(5/6)^2 = 75/216$
$p(2) = (3\ 2)(1/6)^2(5/6)^1 = 15/216$
$p(3) = (3\ 3)(1/6)^3(5/6)^0 = 1/216$

Ex. Screws produced by a certain company will be defective with probability .01 independently of each other. If the screws are sold in packages of 10, what is the probability that two or more screws are defective?

X = number of defective screws.
X is binomial(10,0.01)

$P(X \geq 2) = 1 - P(X<2) = 1 - P(X=0) - P(X=1) = 1 - (10\ 0)(.01)^0(0.99)^{10} - (10\ 1)(.01)^1(0.99)^9 = .004$

Expected value of a binomial(n,p) r.v.

Use linearity of expectation: E(X)=Np

Variance of a binomial(n,p) r.v.

Use rule for adding variances of i.i.d. r.v.'s:

Var(X) = Np(1-p)

Properties of a binomial random variable:

Ex. Sums of two independent Binomial random variables.

X is binomial(n,p)
Y is binomial(m,p)
Z=X+Y. Use convolution formula:

$$p_Z(z) = \sum_{k=0}^{n} p_X(k) p_Y(z-k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \binom{m}{z-k} p^{z-k} (1-p)^{m-z+k}$$

$$= p^z (1-p)^{n+m-z} \sum_{k=0}^{n} \binom{n}{k} \binom{m}{z-k} = \binom{n+m}{z} p^z (1-p)^{n+m-z}$$

Z is binomial(n+m,p)

Can also be proved using mgf's:
M_X(t)=(pexp(t)+1-p)^n
M_Y(t)=(pexp(t)+1-p)^m
Now use the uniqueness theorem for m.g.f.'s.

Multinomial r.v.'s

In the binomial case, we counted the number of outcomes in a binary experiment (the number of heads and tails). The multinomial case generalizes this from the binary to the k-ary case.

Let n_i=number of observed events in i-th slot; p_i = prob of falling in i-th slot.
P(n_1,n_2,…,n_k)=(N!/ n_1!*n_2!*…*n_k!)*p_1^n_1*p_2^n_2*…*p_k^n_k.
The first term is the usual multinomial combinatoric term (familiar from chapter 1); the second is the probability of observing any sequence with n_1 1's, n_2 2's, …, and n_k k's.

To compute the moments here, just note that n_i is itself Bin(N,p_i). So E(n_i)=Np_i and V(n_i)=Np_i(1-p_i).

For the covariances, note that n_i+n_j is Bin(N,p_i+p_j). So
V(n_i+n_j)= N(p_i+p_j)(1-p_i-p_j).
But we also know that
V(n_i+n_j)=V(n_i)+ V(n_j)+2C(n_i,n_j). Put these together and we get that
C(n_i,n_j)=-Np_ip_j.
Note that it makes sense that the covariance is negative: since N is limited, the more of n_i we see, the fewer n_j we'll tend to see.

The Poisson Random Variable

A random variable X, taking on the values 0, 1, 2, ….., is said to be a Poisson random variable with parameter λ if for some λ > 0,
p(i) = P(X=i) = exp(-λ)λ^i /i! for i=0,1,2,3,……

Valid pmf:

$$\sum_{i=0}^{\infty} p(i) = \sum_{i=0}^{\infty} e^{-\lambda}\frac{\lambda^i}{i!} = e^{-\lambda}\sum_{i=0}^{\infty}\frac{\lambda^i}{i!} = e^{-\lambda}e^{\lambda} = 1$$

The Poisson random variable has a large range of applications. A major reason for this is that a Poisson random variable can be used as an approximation for a binomial random variable with parameter (n,p) when n is large and p is small. *(Rare events)*

Let X be a binomial random variable and let λ=np. (λ moderate)

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

$$= \frac{n(n-1)...(n-i+1)}{i!}\frac{\lambda^i}{n^i}\left(1-\frac{\lambda}{n}\right)^{n-i} = \frac{n(n-1)...(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^i}$$

For n large and λ moderate,

$$\left(1-\frac{\lambda}{n}\right)^n \approx e^{-\lambda}; \qquad \frac{n(n-1)...(n-i+1)}{n^i} \approx 1; \qquad \left(1-\frac{\lambda}{n}\right)^i \approx 1.$$

Hence,

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

Ex. If X and Y are independent Poisson random variables with parameters λ_1 and λ_2 respectively, then Z=X+Y is Poisson with parameters $\lambda_1+\lambda_2$.

Verify for yourself, using either m.g.f. (easiest) or convolution formula.

Rare event

Ex. A typesetter, on the average makes one error in every 500 words typeset. A typical page contains 300 words. What is the probability that there will be no more than two errors in five pages?

Assume that each word is a Bernoulli trial with probability of success 1/500 and that the trials are independent.

X = number of errors in five pages (1500 words)
X is binomial(1500,1/500)

$$P(X \le 2) = \sum_{x=0}^{2} \binom{1500}{x} \left(\frac{1}{500}\right)^{x} \left(\frac{499}{500}\right)^{1500-x} = 0.4230$$

Use the Poisson approximation with λ=np=1500/500=3. X is approximately Poisson(3).

$$P(X \le 2) \approx e^{-3} + 3e^{-3} + \frac{3^2 e^{-3}}{2} = 0.4232$$

If X is a Poisson random variable with parameter λ, then E(X)= λ and Var(X)= λ.

$$E(X) = \sum_{i=0}^{\infty} i e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} \lambda \sum_{i=1}^{\infty} \frac{\lambda^{i-1}}{(i-1)!} = e^{-\lambda} \lambda e^{\lambda} = \lambda$$

$$E(X^2) = \sum_{i=0}^{\infty} i^2 e^{-\lambda} \frac{\lambda^i}{i!} = \lambda \sum_{i=1}^{\infty} i \frac{e^{-\lambda} \lambda^{i-1}}{(i-1)!} = \lambda \sum_{ji=0}^{\infty} (j+1) \frac{e^{-\lambda} \lambda^j}{j!}$$

$$= \lambda\left(\sum_{j=0}^{\infty} j \frac{e^{-\lambda} \lambda^j}{j!} + \sum_{j=0}^{\infty} \frac{e^{-\lambda} \lambda^j}{j!}\right) = \lambda(E(X)+1) = \lambda(\lambda+1)$$

Thus, $Var(X) = \lambda(\lambda+1) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

(Compare with Binomial: Np(1-p), N large, p small -> Np.)

MGF:

$E(e^{sX}) = \sum_{i=0}^{\infty} e^{si} e^{-\lambda} \frac{\lambda^i}{i!} = e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda e^s)^i}{i!} = e^{-\lambda} e^{\lambda e^s} = e^{\lambda(e^s - 1)}$; note that this is the limit of the binomial case, (1+(exp(t)-1)p)^n.

Poisson Process

A Poisson process is a model for counting occurrences, or events, over an interval. For example, we are often interested in situations where events occur at certain points in time. *(Time is continuous)*

Ex. The number of customers' entering a post office on a given day.

How many customers enter a store in t hours?

Assume that the customers arrive at random time points.

- Split the t hours into n intervals of very small length *(say 1 s)*.
- *If λ customers arrive on average during an hour, then approximately λt customers will arrive during t hours.* The probability that a customer arrives during each interval is λt/n.
- Only one customer can arrive during each time interval.
- Each person arrives independently.

The intervals can be represented as a sequence of n independent Bernoulli trials with probability of success λt/n in each. *Use Poisson approximation (n large, p small).*

A Poisson process having rate λ means that the number of events occurring in any fixed interval of length t units is a Poisson random variable with mean λt. The value λ is the rate per unit time at which events occur and must be empirically determined. We write that the number of occurrences during t time units as N(t). N(t) is Poison(λt).

Ex. Customers enter a post office at the rate of 1 every 5 minutes.

What is the probability that no one enters during a 10 minute time period?

Let N(t) = the number of customers entering the post office during time t.
N(t) is Poisson(λt) where λ=1 and t=2.
P(N(t) = 0) = exp(-λt) = exp(-2) = 0.135

What is the probability that at least two people enter during a 10 minute time period?

P(X>=2) = 1- P(X<2) = 1-(P(N(2) = 0) + P(N(2) = 1)) = 1-(exp(-2) + 2exp(-2)) = 1-(0.135 + 2*0.135) = 0.595

Geometric distribution

Consider that n independent Bernoulli trials are performed. Each of these trials has probability p of success and probability (1-p) of failure.

Let X = number of unsuccessful trials preceding the first success.

X is a discrete random variable that can take on values of 0,1, 2, 3,.....

S, FS, FFS, FFFS, FFFFS, FFFFFS, FFFFFFS, ...

P(X=0) = p
P(X=1) = (1-p)p
P(X=2) = $(1\text{-}p)^2$p
....
P(X=n) = $(1\text{-}p)^n$p
....

A random variable is called a Geometric(p) random variable if it has the pmf,
p(k) = $(1\text{-}p)^k$p for k=0,1,2,…..

Valid pmf:
sum_{k=0}^inf p$(1\text{-}p)^k$ = p /(1-(1-p)) = p/p = 1

Ex. A fair coin is flipped until a head appears. What is the probability that the coin needs to be flipped more than 5 times?

X = number of unsuccessful trials preceding the first H.

$$P(X \geq 5) = 1 - P(X < 5) = 1 - p\sum_{j=0}^{4}(1-p)^j = 1 - p\frac{1-(1-p)^5}{1-(1-p)}$$

$$= 1 - p\frac{1-(1-p)^5}{p} = (1-p)^5 = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

This is the same probability as getting 5 straight tails. (Note that the combinatoric term here is 5 choose 5, i.e., 1.)

Expected value

Write q=1-p.

$$E(X)=\sum_{n=0}^{\infty} nq^n p = pq\sum_{n=0}\frac{d}{dq}(q^n) = pq\frac{d}{dq}(\sum_{n=0}^{\infty} q^n) = pq\frac{d}{dq}(\frac{1}{1-q}) = \frac{pq}{(1-q)^2} = \frac{pq}{p^2} = \frac{q}{p}.$$

Converging power series. $\frac{d}{dq}(\sum_{n=0}^{\infty} a_n q^n) = \sum_{n=0}^{\infty}\frac{d}{dq}(a_n q^n)$

Variance

Use V(X)=E(X^2)-E(X)^2, and use mgf to get E(X^2).

$M_X(t)=\sum_{n=0}^{\infty} e^{tn}q^n p = \sum_{n=0}^{\infty}(qe^t)^n p = \frac{p}{1-qe^t}$, for qexp(t)<1; otherwise infinite. Take second derivative; find V(X)=q/(p^2).

Negative Binomial

Consider that n independent Bernoulli trials are performed. Each of these trials has probability p of success and probability (1-p) of failure.

X = number of failures preceding the r-th success.

X takes values 0,1,2,…

$P(X=0) = p^r$	SSSSSSS
$P(X=1) = (r\ 1)p^r(1-p)$	SSSSFSSS
$P(X=2) = (r+1\ 2)p^r(1-p)^2$	SSSSFSFSS
$P(X=3) = (r+2\ 3)p^r(1-p)^3$	SSSSFFSFSS

…….

$P(X=n) = (r+n-1\ x)p^r(1-p)^n$

This random variable is called a Negative binomial(r,p) random variable.

Valid pdf: Need to use negative binomials.

The negative binomial distribution is used when the number of successes is fixed and we're interested in the number of failures before reaching the fixed number of successes.

If X is NegBin(r,p), then X=sum_{i=1} ^r X_i, with X_i independent Geo(p) r.v.'s.
E(X)=rq/p; V(X)=rq/(p^2).

MGF of Geo(p):

E(exp(tX_i))= $\sum_{n=0}^{\infty} e^{tn}q^n p = p\sum_{n=0}^{\infty}(qe^t)^n = p\sum_{n=0}^{\infty}(qe^t)^n = p/(1-qe^t)$ for qexp(t)<1.

MGF of NegBin(r,p)=[p/(1-qexp(t))]^r.

<u>Ex.</u> Find the expected number of times one needs to roll a dice before getting 4 sixes.

X = number of rolls before getting 4 sixes.
X is negative binomial with r=4 and p=1/6.
E(X) = r/p = 4/(1/6) = 24.

Hypergeometric random variables

Assume we have a box that contains m red balls and (N-m) white balls. Suppose we choose n different balls from the box, without replacement.

Let X = number of red balls.
X takes value from 0 to n.

P(X=0) = (m 0)(N-m n)/(N n)
P(X=1) = (m 1)(N-m n-1)/(N n)

P(X=k) = (m k)(N-m n-k)/(N n)

A random variable is called a Hypergeometric(n,N,m) random variable if it has the pmf,
P(X=k) = (m k)(N-m n-k)/(N n) for k = 0, 1, …. n

Show that this is a valid pmf.

Suppose that we have a population of size N that consists of two types of objects. For example, we could have balls in an urn that are *red* or *green*, a population of people who are either *male* or *female*, etc. Assume there are *m* objects of type 1, and N-m objects of type 2.

Let X = number of objects of type 1 in a sample of n objects.
X is Hypergeometric with parameters (n,N,m).

Ex. Suppose 25 screws are made, of which 6 are defective. Suppose we randomly sample 10 screws and place them in a package. What is the probability that the package contains no defective screws?

X = number of defective screws in the package.

X is hypergeometric with n=10, N=25 and m=6.

P(X=0) = (6 0)(19 10)/(25 10) = 0.028

Expected value

Need: $x\binom{m}{x} = m\binom{m-1}{x-1}$ and $\binom{N}{n} = \frac{N}{n}\binom{N-1}{n-1}$.

$$E(X) = \sum_{x=0}^{n} x\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n} \frac{m\binom{m-1}{x}\binom{N-m}{n-x}}{\frac{N}{n}\binom{N-1}{n-1}}$$

$$= \frac{mn}{N}\sum_{x=0}^{n} \frac{\binom{m-1}{x}\binom{N-m}{n-x}}{\binom{N-1}{n-1}} = \frac{mn}{N}$$

Variance

$$Var(X) = \frac{mn}{N}\left(\frac{(N-m)(N-n)}{N(N-1)}\right)$$

As population size N becomes large, hypergeometric probabilities converge to binomial probabilities (sampling without replacement -> sampling with replacement). To see this, just set m=Np (where p=fraction of red balls), and write out

(m k)(N-m n-k)/(N n) = (Np k)(N(1-p) n-k)/(N n)
= (n k)[(Np) (Np-1)… (Np-k+1) (N(1-p)) (N(1-p)-1)… (N(1-p)-n+k+1)/(N)(N-1)…(N-n+1)]
≅(n k) p^k (1-p)^(n-k).

Uniform Random Variables

$$f(x) = \begin{cases} \dfrac{1}{\beta - \alpha} & \alpha < x < \beta \\ 0 \text{ otherwise} \end{cases}.$$

Expected Value and Variance:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx = \frac{1}{\beta - \alpha}\left[\frac{x^2}{2}\right]_{\alpha}^{\beta} = \frac{\beta^2 - \alpha^2}{2(\beta - \alpha)} = \frac{(\beta - \alpha)(\beta + \alpha)}{2(\beta - \alpha)} = \frac{(\beta + \alpha)}{2}.$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{\alpha}^{\beta} \frac{x^2}{\beta - \alpha} dx = \frac{1}{\beta - \alpha}\left[\frac{x^3}{3}\right]_{\alpha}^{\beta} = \frac{\beta^3 - \alpha^3}{3(\beta - \alpha)} = \frac{(\beta - \alpha)(\beta^2 + \alpha\beta + \alpha^2)}{3(\beta - \alpha)}$$

$$= \frac{(\beta^2 + \alpha\beta + \alpha^2)}{3}.$$

$$Var(X) = \frac{(\beta^2 + \alpha\beta + \alpha^2)}{3} - \frac{(\alpha + \beta)^2}{4} = \frac{(\beta - \alpha)^2}{12}$$

Ex. Buses arrive at specified stop at 15-minute intervals starting at 7 AM. (7:00, 7:15, 7:30, 7:45, ….) If the 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 the bus;
(b) more than 10 minutes for the bus;

X = number of minutes past 7 the passenger arrives.

X is a uniform random variable over the interval (0,30).

(a) The passenger waits less than 5 minutes if he arrives either between 7:10 and 7:15 or between 7:25 and 7:30.

$$P(10 < X < 15) + P(25 < X < 30) = \int_{10}^{15} \frac{1}{30} dx + \int_{25}^{30} \frac{1}{30} dx = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$

(b) The passenger waits more than 10 minutes if he arrives either between 7:00 and 7:05 or between 7:15 and 7:20.

$$P(0 < X < 5) + P(15 < X < 20) = \int_{0}^{5} \frac{1}{30} dx + \int_{15}^{20} \frac{1}{30} dx = \frac{5}{30} + \frac{5}{30} = \frac{1}{3}$$

Exponential Random Variables

X is an exponential random variable with parameters λ (λ>0) if the pdf of X is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if x} \geq 0 \\ 0 & \text{if x} < 0 \end{cases}$$

The cdf is given by

$F(a) = P(X \leq a) = \int_0^a \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x} \right]_0^a = 1 - e^{-\lambda a}$ for a>=0.

The exponential can be used to model lifetimes or the time between unlikely events. They are used to model waiting times, at telephone booths, at the post office and for time until decay of an atom in radioactive decay.

Expected Value and Variance:

Let X be an exponential random variable.

$$E(X) = \int_0^\infty x\lambda e^{-\lambda x} dx = \left[-xe^{-\lambda x} \right]_0^\infty + \int_0^\infty e^{-\lambda x} dx = 0 + \left[-\frac{e^{-\lambda x}}{\lambda} \right]_0^\infty = \frac{1}{\lambda}$$

$$E(X^2) = \int_0^\infty x^2 \lambda e^{-\lambda x} dx = \left[-x^2 e^{-\lambda x} \right]_0^\infty + \int_0^\infty 2xe^{-\lambda x} dx = 0 + \frac{2}{\lambda} E(X) = \frac{2}{\lambda^2}$$

$Var(X) = E(X^2) - (E(X))^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$; thus the standard deviation (the "scale" of the density) is 1/λ.

These formulas are easy to remember if you keep in mind the following fact: if X ~ exp(1) and Y=X/λ, c>0, then Y ~ exp(λ). Thus λ is really just a scaling factor for Y.

Ex. Suppose that the length of a phone call in minutes is an exponential random variable with λ=1/10. If somebody arrives immediately ahead of you at a public telephone booth, find the probability you have to wait

(a) more than 10 minutes;
(b) between 10 and 20 minutes.

Let X = length of call made by person in the booth.

(a) $P(X > 10) = 1 - P(X \leq 10) = 1 - F(10) = 1 - (1 - e^{-10/10}) = e^{-1} \approx 0.368$

(b) $P(10 < X < 20) = F(20) - F(10) = (1 - e^{-20/10}) - (1 - e^{-10/10}) = e^{-1} - e^{-2} \approx 0.233$

The *Memoryless* Property of the Exponential.

Definition: A nonnegative random variable is memoryless if
$P(X > s + t \mid X > s) = P(X > t)$ for all s,t>0.

P(X>s+t | X>s) = (1-F(s+t))/(1-F(s)). Defining G(x)=1-F(x), the memoryless property may be restated as
G(s+t)=G(s)G(t) for all s,t>0.

We know that G(x) must be non-increasing as a function of x, and 0<G(x)<1. It is known that the only such continuous function is

$$G(x) = \begin{cases} 1, & x<0 \\ \exp(-\lambda x), & x>0, \end{cases}$$

which corresponds to the exponential distribution of rate λ.

Thus we have proved:

Fact 1: Exponential random variables are memoryless.
Fact 2: The Exponential distribution is the only continuous distribution that is memoryless.

There is a similar definition of the memoryless property for discrete r.v.'s; in this case, the geometric distribution is the only discrete distribution that is memoryless (the proof is similar).

Ex. Suppose the lifetime of a light bulb is exponential with λ=1/1000. If the light survives 500 hours what is the probability it will last another 1000 hours.

Let X = the lifetime of the bulb
X is exponential with λ=1/1000.

$$P(X > 1500 \mid X > 500) = P(X > 1000) = \int_{1000}^{\infty} \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_{1000}^{\infty} = e^{-1000\lambda} = e^{-1000/1000} = e^{-1}$$

Derivation of the exponential pdf: think of the Poisson process:
N=L/dx i.i.d. Bernoulli variables of parameter p=λdx. What is P(gap of length x between events)? p(1-p)^(x/dx) = λdx (1-λdx)^(x/dx). Now let dx -> 0.

λdx (1-λdx)^(x/dx) ≈ λdx exp(-λx). So the pdf of the gap length is λexp(-λx).

The Gamma Random Variable

X is a Gamma random variable with parameters (α,λ) (α, λ > 0) if the pdf of X is given by

$$f(x) = \begin{cases} \dfrac{\lambda e^{-\lambda x}(\lambda \mathrm{x})^{\alpha\text{-}1}}{\Gamma(\alpha)} & \text{if x} \geq 0 \\ 0 & \text{if x} < 0 \end{cases}$$

where $\Gamma(\alpha) = \int_0^\infty e^{-y} y^{\alpha-1} dy$. (The gamma function)

It can be shown, through integration by parts, that $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$. *(Verify at home)*
Also a simple calculation shows that: $\Gamma(1) = 1$.

Combining these two facts gives us for integer-valued α (α=n),
$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = \cdots = (n-1)(n-2)\cdots 2\Gamma(1) = (n-1)!$.

Additionally it can be shown that $\Gamma(1/2) = \sqrt{\pi}$. *(Verify at home)*

Expected Value and Variance:

$$E(X) = \frac{1}{\Gamma(\alpha)}\int_0^\infty \lambda x e^{-\lambda x}(\lambda x)^{\alpha-1} dx = \frac{1}{\Gamma(\alpha)\lambda}\int_0^\infty \lambda e^{-\lambda x}(\lambda x)^{\alpha} dx = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\lambda} = \frac{\alpha}{\lambda}$$

$$E(X^2) = \frac{1}{\Gamma(\alpha)}\int_0^\infty \lambda x^2 e^{-\lambda x}(\lambda x)^{\alpha-1} dx = \frac{1}{\Gamma(\alpha)\lambda^2}\int_0^\infty \lambda e^{-\lambda x}(\lambda x)^{\alpha+1} dx = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\lambda^2} = \frac{(\alpha+1)\alpha}{\lambda^2}$$

$$Var(X) = E(X^2) - (E(X))^2 = \frac{(\alpha+1)\alpha}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}$$

Many random variables can be seen as special cases of the gamma random variable.

(a) If X is gamma with α=1, then X is an exponential random variable.
(b) If X is gamma with λ=1/2 and α=n/2 is χ^2 with n degrees of freedom. (We'll discuss this soon.

<u>Ex.</u> Let X be a Gamma distribution with parameter (α, λ).

$$f(x) = \frac{1}{\Gamma(\alpha)}\lambda e^{-\lambda x}(\lambda x)^{\alpha-1} \text{ for } 0<=x<\infty.$$

Calculate $M_X(t)$:

$$M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx}\frac{1}{\Gamma(\alpha)}\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}dy = \int_0^\infty \frac{1}{\Gamma(\alpha)}\lambda e^{-(\lambda-t)x}(\lambda x)^{\alpha-1}dx$$

$$= \frac{\lambda^\alpha}{(\lambda-t)^\alpha}\int_0^\infty \frac{1}{\Gamma(\alpha)}(\lambda-t)e^{-(\lambda-t)x}((\lambda-t)x)^{\alpha-1}dx = \frac{\lambda^\alpha}{(\lambda-t)^\alpha} = \left(\frac{\lambda}{\lambda-t}\right)^\alpha$$

$$M_X{}'(t) = \alpha\left(\frac{\lambda}{\lambda-t}\right)^{\alpha-1}\frac{\lambda}{(\lambda-t)^2} = \frac{\alpha}{\lambda}\left(\frac{\lambda}{\lambda-t}\right)^{\alpha+1}$$

$$E(X) = M_X{}'(0) = \frac{\alpha}{\lambda}\left(\frac{\lambda}{\lambda-0}\right)^{\alpha+1} = \frac{\alpha}{\lambda}$$

$$M_X{}''(t) = \frac{\alpha(\alpha+1)}{\lambda}\left(\frac{\lambda}{\lambda-t}\right)^{\alpha+1}\frac{\lambda}{(\lambda-t)^2} = \frac{\alpha(\alpha+1)}{\lambda^2}\left(\frac{\lambda}{\lambda-t}\right)^{\alpha+3}$$

$$E(X^2) = M_X{}''(0) = \frac{\alpha(\alpha+1)}{\lambda^2}\left(\frac{\lambda}{\lambda-0}\right)^{\alpha+3} = \frac{\alpha(\alpha+1)}{\lambda^2}$$

Let X and Y be independent gamma random variables with parameters (s,λ) and (t,λ). Calculate the distribution of Z=X+Y.

$$M_{X+Y}(t) = M_X(t)M_Y(t) = \left(\frac{\lambda}{\lambda+t}\right)^{s}\left(\frac{\lambda}{\lambda+t}\right)^{t} = \left(\frac{\lambda}{\lambda+t}\right)^{s+t}$$

Z=X+Y is gamma with parameters (s+t,λ). Can also use convolution:

$$f_z(z) = \int_{-\infty}^{\infty} f_X(z-y)f_y(y)dy = \int_{-\infty}^{\infty}\frac{1}{\Gamma(s)}\lambda e^{-\lambda(z-y)}(\lambda(z-y))^{s-1}\frac{1}{\Gamma(t)}\lambda e^{-\lambda y}(\lambda y)^{t-1}dy$$

$$= \frac{1}{\Gamma(s)\Gamma(t)}\lambda^{s+t}e^{-\lambda z}\int_{-\infty}^{\infty}(z-y)^{s-1}y^{t-1}dy$$

$$\int_{-\infty}^{\infty}(z-y)^{s-1}y^{t-1}dy = \frac{z^{s+t-2}}{z^{s+t-2}}\int_{-\infty}^{\infty}(z-y)^{s-1}y^{t-1}dy = z^{s+t-2}\int_{-\infty}^{\infty}(1-\frac{y}{z})^{s-1}(\frac{y}{z})^{t-1}dy$$

$$= z^{s+t-2}z\int_{-\infty}^{\infty}(1-x)^{s-1}(x)^{t-1}dx = z^{s+t-1}B(s,t) = z^{s+t-1}\frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

$$f_z(z) = \frac{1}{\Gamma(s)\Gamma(t)}\lambda^{s+t}e^{-\lambda z}\int_{-\infty}^{\infty}(z-y)^{s-1}y^{t-1}dy = \frac{1}{\Gamma(s)\Gamma(t)}\lambda^{s+t}e^{-\lambda z}z^{s+t-1}\frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)}$$

$$= \frac{1}{\Gamma(s+t)}\lambda e^{-\lambda z}(\lambda z)^{s+t-1}$$

Z is gamma with parameters (s+t,λ)

In general, we have the following proposition:

Proposition: If X_i, i=1,….n are independent gamma random variables with parameters (t_i,λ) respectively, then $\sum_{i=1}^{n} X_i$ is gamma with parameters $(\sum_{i=1}^{n} t_i, \lambda)$.

Proof: Use the example above and prove by induction.

Ex. Sums of independent exponential random variables. Exponential with parameter λ is equivalent to Gamma (1,λ). It follows that if X_i, i=1,….n are independent exponential random variables with parameter λ, then $\sum_{i=1}^{n} X_i$ is gamma with parameters (n, λ).

The Beta Random Variable

X is a Beta random variable with parameters (a,b) if the pdf of X is given by

$$f(x) = \begin{cases} \dfrac{1}{B(a,\mathrm{b})} x^{a-1}(1 - \mathrm{x})^{\mathrm{b-1}} & \text{if } 0 < \mathrm{x} < 1 \\ 0 & o\text{therwise} \end{cases}$$

where $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. (The Beta function)

The Beta family is a flexible way to model random variables on the interval [0,1]. It is often used to model proportions (e.g., the prior distribution of the fairness of a coin).

The following relationship exists between the beta and gamma function:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \qquad (\textit{Verify: } \Gamma(a)\Gamma(b) = \ldots = \Gamma(a+b)B(a,b))$$

Expected Value and Variance:

$$E(X) = \frac{1}{B(a,b)}\int_0^1 xx^{a-1}(1-x)^{b-1}dx = \frac{1}{B(a,b)}\int_0^1 x^{a}(1-x)^{b-1}dx = \frac{B(a+1,b)}{B(a,b)}$$

$$= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} = \frac{a\Gamma(a)\Gamma(a+b)}{(a+b)\Gamma(a+b)\Gamma(a+b)} = \frac{a}{(a+b)}$$

$$Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

We will see an application of the beta distribution to order statistics of samples drawn from the uniform distribution soon.

Normal Random Variables

X is a normal random variable with parameters μ and σ^2 if the pdf of X is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad \text{for x in } (-\infty, \infty).$$

This is the density function for the so-called Bell curve. The normal distribution shows up frequently in Probability and Statistics, due in large part to the Central Limit Theorem.

If X is a normally distributed random variable with μ=0 and σ=1, then X has a standard normal distribution. The pdf of X is then given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\mu^2}{2}} \quad \text{for x in } (-\infty, \infty).$$

f(x) is a valid pdf, because f(x)>=0 and $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} dx = 1$

This last result is not immediately clear. See book for proof. The idea is to compute the joint pdf of two iid Gaussians, for which the normalization factor turns out to be easier to compute.

Fact: If X is normally distributed with parameters μ and σ^2, then Y=aX+b is normally distributed with parameters aμ+b and $a^2\sigma^2$.

Proof:

$$F_Y(y) = P(Y \le y) = P(aX + b \le y) = P(X \le \frac{y-b}{a}) = F_X(\frac{y-b}{a})$$

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(\frac{y-b}{a}) = \frac{1}{a} f_X(\frac{y-b}{a}) = \frac{1}{\sqrt{2\pi a\sigma}} e^{-(\frac{y-b}{a}-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi a\sigma}} e^{-(y-(b+a\mu))^2/2a^2\sigma^2}$$

Y is normally distributed with parameters aμ+b and $a^2\sigma^2$.

An important consequence of this fact is that if X is normally distributed with parameters μ and σ^2, then $Z = \frac{X-\mu}{\sigma}$ is standard normal.

Expected Value and Variance:

Let Z be a standard normal random variable.

$$E(Z) = \int_{-\infty}^{\infty} zf(z)dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ze^{-z^2/2}dx = \frac{1}{\sqrt{2\pi}}\left[e^{-z^2/2}\right]_{-\infty}^{\infty} = 0$$

V(Z)=1; this is easiest to derive from the mgf, which we'll derive below.

Let Z be a standard normal random variable and let X be normally distributed with parameters μ and σ^2.

$Z = \frac{X-\mu}{\sigma}$ and hence $X = \sigma Z + \mu$.

$$E(X) = E(\mu + \sigma Z) = \mu + \sigma E(Z) = \mu$$

$$Var(X) = Var(\mu + \sigma Z) = \sigma^2 Var(Z) = \sigma^2$$

It is traditional to denote the cdf of a standard normal random variable Z by Φ(z).

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-y^2/2}dy$$

The values of Φ(z) for non-negative z are given in Table 5.1.

For negative values of z use the following relationship:

$$\Phi(-z) = P(X \le -z) = P(Z > z) = 1 - P(Z \le z) = 1 - \Phi(z).$$

Also if X is normally distributed with parameters μ and σ^2 then,

$$F_X(x) = P(X \le x) = P(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}) = P(Z \le \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma}).$$

Ex. If X is a normally distributed with parameters μ=3 and σ^2=9, find

(a) P(2<X<5)
(b) P(X>0)
(c) P(|X-3)>6)

$$P(2 < X < 5) = P(\frac{2-3}{3} < \frac{X-3}{3} < \frac{5-3}{3}) = P(-\frac{1}{3} < Z < \frac{2}{3}) = \Phi(\frac{2}{3}) - \Phi(-\frac{1}{3})$$

$$= \Phi(\frac{2}{3}) - (1 - \Phi(\frac{1}{3})) = .7454 - (1 - .6293) = .3779$$

$$P(X > 0) = P(\frac{X-3}{3} > \frac{0-3}{3}) = P(Z > -1) = 1 - \Phi(-1) = 1 - (1 - \Phi(1)) = \Phi(1) = .8413$$

$$P(|X-3| > 6) = P(X > 9) + P(X < -3) = P(\frac{X-3}{3} > \frac{9-3}{3}) + P(\frac{X-3}{3} < \frac{-3-3}{3}) = (1 - \Phi(2)) + \Phi$$

$$= (1 - \Phi(2)) + (1 - \Phi(2)) = 2(1 - \Phi(2)) = 0.0456$$

Ex. MGF of a Normal random variable.

Let Z be a standard normal random variable:

$$M_Z(t) = E(e^{tZ}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz} e^{-z^2/2} dx$$

$$e^{tz} e^{-z^2/2} = e^{-(z^2-2tz)/2} = e^{-(z^2-2tz-t^2+t^2)/2} = e^{t^2/2} e^{-(z^2-2tz+t^2)/2} = e^{t^2/2} e^{-(z-t)^2/2}$$

$$M_Z(t) = E(e^{tZ}) = e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} dx = e^{t^2/2}$$

To obtain the mgf for an arbitrary normal random variable, use the fact that if X is a normal random variable with parameters (μ, σ^2) and Z is standard normal, then X=μ+Zσ.

$$M_X(t) = E(e^{tX}) = E(e^{t(\mu+Z\sigma)}) = E(e^{t\mu} e^{tZ\sigma}) = e^{t\mu} E(e^{(t\sigma)Z}) = e^{t\mu} M_Z(t\sigma)$$

$$= e^{t\mu} e^{(t\sigma)^2/2} = e^{t^2\sigma^2/2 + t\mu}$$

Proposition: If X and Y are independent random variables with parameters (μ_1, σ^2_1) and (μ_2, σ^2_2) respectively, then X+Y is normal with mean $\mu_1+\mu_2$ and variance $\sigma^2_1+\sigma^2_2$.

Proof:

$$M_{X+Y}(t) = M_X(t)M_Y(t) = e^{t^2\sigma_1^2/2+t\mu_1}e^{t^2\sigma_2^2/2+t\mu_2} = e^{t^2(\sigma_1^2+\sigma_2^2)/2+t(\mu_1+\mu_2)}$$

Hence X+Y is normal with mean $\mu_1+\mu_2$ and variance $\sigma^2_1+\sigma^2_2$.

Ex. Let X and Y be independent standard normal random variables, and let Z=X+Y. Calculate the pdf of Z.

$$f_z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-(z-x)^2/2}dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}e^{-(x^2+(z-x)^2)/2}dx$$

$$x^2+(z-x)^2 = x^2+(z^2-2xz+x^2) = z^2-2xz+2x^2 = 2(x^2-xz+\frac{z^2}{4})+\frac{z^2}{2} = 2(x-\frac{z}{2})^2+\frac{z^2}{2}$$

$$f_z(z) = \frac{1}{2\pi}e^{-z^2/(2*2)}\int_{-\infty}^{\infty}e^{-(x-z/2)^2}dx = \frac{1}{2\pi}e^{-z^2/(2*2)}\sqrt{2\pi}\frac{1}{1/\sqrt{2}} = \frac{1}{\sqrt{2\pi}\sqrt{2}}e^{-z^2/(2(\sqrt{2})^2)}$$

Z is normally distributed with parameters 0 and $\sqrt{2}$.

<u>Ex.</u> (Chi-square random variables)

Let Z be a standard normal random variable. Let Y= Z^2.

Recall that if Z is a continuous random variable with probability density f_Z then the distribution of Y= Z^2 is obtained as follows: $f_Y(y) = \frac{1}{2\sqrt{y}}[f_Z(\sqrt{y}) + f_Z(-\sqrt{y})]$.

(Ex. 7b p. 223)

Also, $f_z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$

$$f_Y(y) = \frac{1}{2\sqrt{y}}[f_Z(\sqrt{y}) + f_Z(-\sqrt{y})] = \frac{1}{2\sqrt{y}} \frac{2}{\sqrt{2\pi}} e^{-y/2} = \frac{\frac{1}{2} e^{-y/2} (y/2)^{1/2-1}}{\sqrt{\pi}}$$

Since $\sqrt{\pi} = \Gamma(1/2)$, Y is gamma with parameters (1/2,1/2).

<u>Proposition:</u> Let Z_i, i=1,….n be independent standard normal random variables, then $\sum_{i=1}^{n} Z_i^2$ is gamma with parameter (n/2,1/2) or χ^2 with n degree of freedom.

Very important in statistical analysis.

Multivariate normal

$P(X) = (2pi)^{-d/2} |C|^{-1/2} \exp[-.5 (X-m)' C^{-1} (X-m)]$, with d=dim(X).
Contours are ellipses, centered at m.

X may be represented as
X=Rz+m,
where RR'=C and z~N(0,I) (i.e., z is iid standard normal); to see this, just apply our change-of-measure formula.

From this representation, we may derive the following key facts:
E(X_i)=m_i;
C(X_i,X_j)=C_ij.
AX=ARz+Am, so AX~N(Am,ACA') – ie, linear transformations preserve Gaussianity. This also implies that marginalization preserves Gaussianity.

Eigenvalues and eigenvectors ("principal components") of C are key.

It's also not hard to prove that conditioning preserves Gaussianity, in the following sense:
Let (X Y) ~ N[(a b),(A C;C' B)]. Then
$X|Y=y \sim N(a+B^{-1}(y-b), A-C B^{-1}C')$.
Three important things to remember here:

1) E(X|y) is linear in y.
2) C(X|y) is smaller than C(X), in the sense that C(X)-C(X|y) is positive semidefinite. Also, C(X|y) is independent of y.
3) The larger the normalized correlation $CB^{-1/2}$, the larger the effect of conditioning. If C=0, then X and Y are independent and conditioning has no effect.

Mixture models

One final important class of probability distributions is the class of densities known as "mixture models." The idea here is that often a single distribution in the families we've discussed so far is inadequate to describe the data. For example, instead of one bell-shaped curve (a single "bump"), our data may be better described by several bumps, i.e.,

f_X(x) = sum_i a_i g_i(x),

where a_i>0, sum a_i =1, and each individual g_i(x) is a well-defined pdf (eg a Gaussian with mean m_i and variance v_i). This gives us much more flexibility in modeling the true observed data.

The a_i's have a natural probabilistic interpretation here. The idea is that X may be modeled as coming from some joint distribution (X,L), where L is a discrete "label" variable and p(L=i)=a_i and f_X(x|L=i) = g_i(x). Thus the idea is that, to generate the data X, we choose a label from the discrete distribution p(L=i)=a_i, then choose X according to g_i(x), and then forget the identity of the label L (i.e., marginalize over L, which corresponds to computing
f_X(x) = sum_i P(X,L=i) = sum_i P(L=i) f_X(x|L=i) = sum_i a_i g_i(x).