Background on Basic Probability Theory on Independent Random Variables

(Students should know at least what are in the text boxes)

1. Distribution Law of Integral-Valued Random Variables

An integral-valued random valuable X can take its value from the set $\{..., -1, 0, +1,...\}$, its distribution law is completely determined if the probability $P(X=n) = p(n)$ for very integer n between $-\infty$ and $+\infty$ is known. In this review, we mainly focus on binary Random Variable, for which: $p(0)=q=1-p$, $p(1)=p$, and $p(n)=0$ for $n\neq 0$ or 1, where $0 \leq p, q \leq 1$ are the conditional binomial parameters.

Probability must satisfy normalization condition $\Sigma p(n) = 1$ (summing over all possible integers). The expectation of $f(X)$ is given by $Ef(X) = \Sigma f(n)\, p(n)$, in particular, the mean is $m = EX = \Sigma n\, p(n)$ and the dispersion or variance is $D=\sigma^2 = E(X-m)^2 = \Sigma (n-m)^2\, p(n)$.

For binary variable, $m = 0*q + 1*p = p$ and $D=\sigma^2 = (0-p)^2\, q + (1-p)^2\, p = p^2\, q + q^2\, p = pq$. (1)

2. **Characteristic Function** (Paul Lévy)

If $f(X) = e^{isX}$, ($i=\sqrt{-1}$, s is a real parameter), its expectation

$$\varphi(s) = Ee^{isX} = \Sigma\, e^{isn}\, p(n) \tag{2}$$

is called characteristic function of X. The Fourier transform of $\varphi(s)$ is obtained by multiplying by e^{-isk} and integrating,

$$p(k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ds\varphi(s)e^{-isk} \tag{3}$$

This inversion formula shows that the distribution law $p(n)$ is completely determined by its characteristic function $\varphi(s)$ (students may read reference http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)).

Here are some nice properties for the characteristic function: $\varphi(0) = 1$, $|\varphi(s)| \leq 1$,

$$m = EX = \Sigma n\, p(n) = -i\left(\frac{d\varphi}{ds}\right)_{s=0} = -i\left(\frac{d(\ln\varphi)}{ds}\right)_{s=0} \tag{4}$$

$$D=\sigma^2 = E(X-m)^2 = \Sigma (n-m)^2\, p(n) = -\left(\frac{d^2(\ln\varphi)}{ds^2}\right)_{s=0} \tag{5}$$

and

For binary variable,

$$\varphi(s) = q\, e^{is0} + p\, e^{is1} = q + p\, e^{is}, \tag{6}$$

$$m = -i\frac{d(\ln\varphi)}{ds}\bigg|_{s=0} = \frac{pe^{is}}{q+pe^{is}}\bigg|_{s=0} = p,$$

$$D = -\frac{d^2(\ln\varphi)}{ds^2}\bigg|_{s=0} = \left[\frac{pe^{is}}{q+pe^{is}} - \frac{p^2e^{i2s}}{(q+pe^{is})^2}\right]\bigg|_{s=0} = p - p^2 = pq,$$ agree with (1).

Another nice property is the rule of **composition of distributions** of independent random variables: if X, Y are independent (e.g. X is the outcome of p'-coin and Y is that of p''-coin, assigning value1 to "head" and 0 to "tail"), and $Z = X + Y$, then the characteristic function of Z is simply the product of the characteristic functions for each random variable:

$$\varphi(s) = Ee^{isZ} = Ee^{is(X+Y)} = Ee^{isX}\, Ee^{isY} = \varphi'(s)\, \varphi''(s) \tag{6}$$

Ex1: Consider a sequence of 500 throws of a die. What is the probability that the total is 1500? Direct enumeration would be quite tedious, since each through is independent and obeys the same probability low with characteristic function φ_0, $\varphi(s) = [\varphi_0(s)]^{500}$,

$$p(1500) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ds\varphi(s)e^{-is1500} = \frac{1}{2\pi}\int_{-\pi}^{\pi} dse^{-is1500}\left(\frac{1}{6}\sum_{j=1}^{6} e^{ijs}\right)^{500}$$

The appropriate terms can be selected by expanding the term inside the parenthesis with multinomial expansion.

3. **Random Walk** review of Tossing *N* coins (Bernoulli trials)

Suppose a random walker starts off at origin, the probability with a step to the right is *p* and to the left is *q=1-p*, the total distance from the origin after *N* steps is $S_N = \sum_{i=1}^{N} Y_i$, where $Y_i = +1$ or -1 (this variable *Y* is related to the 0-1 binary variable *X* by *Y=2X-1*). One can easily verify that the characteristic function of *Y* is $\psi(s)=qe^{-is}+pe^{+is}$, its mean $m_Y=EY=p-q$, and variance $\sigma_Y^2=DY=4pq$. Since each step is independent, the characteristic function for S_N is $\psi_N(s) = \psi(s)^N$ according to (6). So

$$P(S_N = M) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ds\psi(s)^N e^{-isM} = \frac{N!}{[(N+M)/2]![(N-M)/2]!} p^{(N+M)/2}q^{(N-M)/2} = \boxed{\binom{N}{K} p^K q^{N-K} = p_b(K,N,p)}$$

where $K=\Sigma X_i=(N+M)/2$ is the total number of steps to the right and $p_b(K,N,p)$ is the famous **Binomial Distribution**.

4. Difference Equation for the Random Walk

Let $P(S_N = M) = p(M,N)$, it obviously satisfies the following difference equation

$p(M, N+1) = qp(M+1, N) + pp(M-1, N)$ with the boundary condition $p(0,0)=1$ and $p(M,0)=0$. If one multiplies e^{iMs} on both sides and sums over *M*, one would get the recursion equation for the characteristic function $\psi_{N+1}(s) = (qe^{-is} + pe^{+is})\psi_N(s) = \psi(s)\psi_N(s)$ which will yield the same solution $\psi_N(s) = \psi(s)^N$ as given by (6).

5. **Binomial to Normal** (N >> 1)

When N is large, Binomial distribution may be approximated by Normal distribution and this can be done by Taylor series expansion of $\ln \psi_N(s)$ about $s=0$:

$$\ln\psi_N(s) = N\ln\psi(s) = N\left(\ln 1 + \left.\frac{d\ln\psi}{ds}\right|_{s=0} s + \frac{1}{2}\left.\frac{d^2\ln\psi}{ds^2}\right|_{s=0} s^2 + \mathrm{O}(s^3)\right) = N\left(ism_Y - \frac{1}{2}s^2\sigma_Y^2 + \cdots\right)$$

$$P(S_N = M) = \frac{1}{\pi}\int_{-\pi/2}^{\pi/2} dse^{-isM+N(ism_Y - s^2\sigma_Y^2/2+\cdots)} \approx \frac{2}{\sqrt{2\pi N\sigma_Y^2}}\exp\left[-\frac{(M-Nm_Y)^2}{2N\sigma_Y^2}\right] = \frac{1}{\sqrt{2\pi Npq}}\exp\left[-\frac{[M-N(p-q)]^2}{8Npq}\right]$$

where $o(s^2)$ is thrown away and the range of integration is extended to $\pm\infty$ because the coefficient of s^2 is very large (linear in *N*). In general, the distribution for sum of *N* independent random variables of any distribution will approach to Normal Distribution (**Central Limit Theory**). Hence Binomial distribution will approach to Normal distribution

$$\boxed{P_b(K,N,p) \approx \frac{1}{\sqrt{2\pi Npq}}\exp\left[-\frac{(K-Np)^2}{2Npq}\right] \quad \text{when N>>1.}} \tag{7}$$

This suggests the normalized variable $Z_N = (K-Np)/\sqrt{Npq} = (K-Nm)/\sqrt{N}\sigma$ (de Moirre, *p=1/2*; Laplace, general *p*; Bernstein $\sqrt{N}$ law), Čebyšev provided precise error bound giving **Law of Large Number** (N→∞):

$$P(|Z_N| \le T) \ge 1 - \frac{1}{T^2} \text{ or } P(|x-m| \le \varepsilon) \ge 1 - \frac{\sigma^2}{N\varepsilon^2} \quad \text{where } x = \frac{K}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i, \varepsilon = T\sqrt{\frac{N}{pq}}, m = p, \sigma^2 = pq. \tag{8}$$

6. **Binomial to Poisson** ($p << 1$, $Np=\lambda$)

If $p << 1$, $\ln\varphi_N(s) = N \ln\varphi(s) = N \ln(q+pe^{is}) = N \ln[1+p(e^{is}-1)] \approx Np(e^{is}-1)] = \lambda (e^{is}-1)$.

$$p_b(K,N,p) = \frac{1}{2\pi}\int_{-\pi}^{\pi} ds\varphi_N(s)e^{-isK} \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} dse^{-isK}e^{-\lambda(1-e^{is})} = e^{-\lambda}\frac{1}{2\pi}\int_{-\pi}^{\pi} dse^{-isK}\left(1+\lambda e^{is}+\cdots+\frac{\lambda^K e^{isK}}{K!}+\cdots\right) = e^{-\lambda}\frac{\lambda^K}{K!} = p_p(K,\lambda)$$

and, for Poisson distribution $p_p(K, \lambda)$, *mean=variance=*λ. (Prof. Serfling told me, the most useful thing to know about such approximation error bound is the inequality $|p_b(K,N,p) - p_p(K, \lambda)| << Np^2$, which gives explicitly the maximum error the Poisson approximation can make regardless).

Ex2. Suppose we have *N=50* DNA sequences, each has length *L=200* base pairs, what is the probability of finding *K=30* sequences that contain the "TATAAA" (TATA-box) word (motif) if each base-pair is independently and identifiably distributed (I.I.D.) according to the uniform law *P(A)=P(C)=P(G)=p(T)=0.25*? Assuming the word can occur in any one of (200-6+1)=195 positions, p=195x$(0.25)^6$=195x2.4414062E-4=0.0476, $\lambda=Np=50x0.0476=2.38$, $p_p(30,2.38)=e^{-2.38}(2.38)^{30}/30!= \exp(-2.38+30*0.867-74.658)=\exp(-51.0)=0.69x10^{-22}$.

7. Large Deviation

Similarly, error bound can be obtained for Normal approximation. Let $K = \sum_{i=1}^{N} X_i$, for $0 < p < a < 1$, Arratia and Gordon (Bull. Of Math, 51:125, 1989) show (please compare this to (8)):

$P(K \geq aN) \leq e^{-NH}$, and

$$P(K = aN + i) \approx \frac{r^i}{\sqrt{2\pi a(1-a)N}}e^{-NH}$$, for *i=0,1,2,…*; or summing all *i*, one has

$$P(K \geq aN) \approx \frac{1}{1-r}\frac{1}{\sqrt{2\pi a(1-a)N}}e^{-NH},$$

where the **relative entropy** $H(a,p) = a\ln\left(\frac{a}{p}\right) + (1-a)\ln\left(\frac{1-a}{1-p}\right)$ and the "odds ratio"

$r(a,p) = \left(\frac{p}{1-p}\right)\Big/\left(\frac{a}{1-a}\right)$ play very important roles. They are related by $\left(\frac{\partial H(a,p)}{\partial a}\right) = -\ln(r)$.

Ex3. What is the probability of 16 or more A's in 20 bp long DNA sequence when P(A)=0.25? From exact calculation p* =P_b(K≥16, N=20,p=0.25) = P_b(K=16,20,0.25) + P_b(K=17,20,0.25) + P_b(K=18,20,0.25) + P_b(K=19,20,0.25) + P_b(K=20, 20, 0.25) = 0.3865x10^{-6}, the main contribution comes from P_b(K=16,20,0.25)=0.3569x10^{-6}. To get normal approximation, $N\sigma^2$=Np(1-p)=10(0.25)(0.75)=3.75, Nm=Np=5, Z=(16-5)/$\sqrt{}$(3.75)=5.68. The chance that a standard normal exceeds this is 0.0069x10^{-6},which is 56 times smaller than the exact value! If one were more careful and included the continuity correction, one would obtain 1-erf((15.5-5)/ $\sqrt{}$(3.75)=0.03038x10^{-6}, which is still 12.7 times too small. Using large deviation theory, a=16/20=0.8, H(0.8,0.25)=0.8ln(3.2)+0.2ln(4/15)=0.93-0.26=0.666. The upper bound the first formula is e^{-NH}=1.64 x10^{-6}(4.24 times too big), the central limit factor is $\sqrt{}$(2π(20)(0.8)(0.2) = $\sqrt{}$(20.11) =4.48. The combination of these two factors is $e^{-NH}/\sqrt{}$(2πNa(1-a) =1.64x10^{-6}/4.48=0.366 x10^{-6}. The odds ratio is (1/3)/(4/1)=1/12, so the correction factor corresponding to "K or more A's" instead of "exactly K A's" is 1/(1-r)=12/11, which is not far from 1. Putting everything together, the large deviation approximation is 0.398x10^{-6}, which is only 3% above the exact value. If a bacterium genome has 1 million base pairs of DNA, the overall statistical significance of observing 16 A's in a 20 bp window is reasonably bounded by 1,000,000p* (more precisely, by (1,000,000-20+1)p*=999,981p*, the number of 20bp windows in the genome). Normal approximation gives 0.0069, over estimated the significance; the large deviation result gives 0.4, so one concludes that even in purely random data, the chance of some 20pb window containing at least 16 A's is not small, and such observation could easily be due chance alone.