BCB BCB/GDCB/STAT/COM S 568 Spring 2011

Homework 2 Solution

February 2, 2011

Exercises

1. **Exercise** Find the probability generating function for the binomial and Poisson distributions. Calculate the mean and variance of both distributions by equations 1.1 and 1.2 in the class notes.

   Solution:

   (1) A random variable X follows the binomial distribution $B(n,p)$ if $\text{Prob}(X=k)=b(k;n,p)$ for $k=0,1,\dots n$, where

   $$b(k;n,p)=\binom{n}{k}p^k q^{n-k},$$

   and $p+q=1$. The associated probability generating function is seen to be

   $$A(s)=\sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}s^k=(ps+q)^n.$$

   We obtain

   $$A'(s)=n(ps+q)^{n-1}p$$

   and

   $$A''(s)=n(n-1)(ps+q)^{n-2}p^2.$$

   Thus the mean is

   $$\mathrm{E}[X]=A'(1)=np,$$

   and the variance is

   $$\begin{aligned}\mathrm{Var}(X) &= A''(1)+A'(1)-A'(1)^2\\ &= n^2p^2-np^2+np-(np)^2\\ &= np-np^2\\ &= npq.\end{aligned}$$

   (2) The Poisson probability distribution with parameter λ is given by the probability assignments

   $$P(k;\lambda)=\frac{\lambda^k}{k!}e^{-\lambda}$$

   for $k=0,1,2,\dots$.

   Thus the associated probability generating function is

   $$A(s)=e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda s)^k}{k!}=e^{-\lambda}e^{\lambda s}=e^{\lambda(s-1)}$$

We get $A'(s) = e^{\lambda(s-1)}\lambda$ and $A''(s) = e^{\lambda(s-1)}\lambda^2$, and thus the mean and variance are calculated as

$$\mathrm{E}[X] = A'(1) = \lambda, \qquad \mathrm{Var}[X] = A''(1) + A'(1) - A'(1)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

2. **Exercise** For a non-negative integer valued random variable X with $\mathrm{Prob}(X = k) = p_k$, rearrange terms in $E[X] = \sum_{k=0}^{\infty} kp_k$ into blocks of terms starting from $k = 1$, $k = 2$, and so forth to see the validity of equation 1.4 in a different way.

   Solution:
   We know that
   $$t_j = p_{j+1} + p_{j+2} + \dots, \qquad j = 0, 1, 2, \dots \quad .$$
   Thus

   $$\begin{array}{lllllll} t_0 = & p_1+ & p_2+ & p_3+ & p_4+ & \dots \\ t_1 = & & p_2+ & p_3+ & p_4+ & \dots \\ t_2 = & & & p_3+ & p_4+ & \dots \\ t_3 = & & & & p_4+ & \dots \end{array}$$

   If we do the summation over all t_j, we will add p_i i times. For example, we add three times for p_3. Therefore, we have:

   $$\begin{aligned} T(1) &= \sum_{k=0}^{\infty} t_k \\ &= \sum_{k=1}^{\infty} kp_k \\ &= 0p_0 + \sum_{k=1}^{\infty} kp_k \\ &= \sum_{k=0}^{\infty} kp_k \\ &= \mathrm{E}[X]. \end{aligned}$$

3. **Exercise** Let X be a random variable that takes on the values 1 and 0 with probabilities p and $q = 1-p$, respectively. Use equation 1.7 to determine the probability distribution of $S = \sum_1^n X_i$, where each of the X_i is independently distributed like X.

   Solution:

   Let X_1,X_2,...,X_n be Bernoulli random variables, the probability distribution is:
   $$f(x) = \begin{cases} p & \text{if } k = 1 \\ q & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$
   The generating function for Bernoulli distribution is:
   $$\begin{aligned} P(s) &= \sum_{k=0}^{\infty} a_k s^k \\ &= a_0 + a_1 s \\ &= p + qs \end{aligned}$$

   From equation 1.7, we can easily see that the generating function for $S = \sum_{i=1}^{n} X_i$ is :
   $$\begin{aligned} P_{x_1+\dots+x_n}(s) &= \prod_{i=1}^{n} P_{x_i}(s) \\ &= \prod_{i=1}^{n} (q + ps) \\ &= (ps + q)^n \end{aligned}$$

So the generating function for S is a binomial generating function (see **Exercise 1**):

$$\text{Prob}\{S = k\} = B(k; n, p) = \binom{n}{k} p^k q^{n-k}$$

We can also extend this method to obtain the probability distribution for the sum of n binomial random variables. Suppose X_1,X_2,...,X_n are binomial random variable, and $X_i \sim B(k; m_i, p)$. Then we get

$$\begin{aligned} P_{x_1+\cdots+x_n}(s) &= \prod_{i=1}^{n} P_{x_i}(s) \\ &= \prod_{i=1}^{n} (ps + 1 - p)^{m_i} \\ &= (ps + 1 - p)^{\sum_{i=1}^{n} m_i} \end{aligned}$$

So the generating function for $S = \sum_{i=1}^{n} X_i$ is also a binomial generating function. And the probability distribution of S follows the binomial distribution:

$$S \sim B(k; \sum_{i=1}^{n} m_i, p)$$

4. **Exercise** Derive equation 1.20 by appropriately summing up the approximate expressions for f_n given in equation 1.19.

   Solution:

$$\begin{aligned} t_n &= \sum_{k=n+1}^{\infty} f_k \\ t_n &\approx \sum_{k=n+1}^{\infty} \left[\frac{(s_1 - 1)(1 - ps_1)}{(r + 1 - rs_1)q} \frac{1}{s_1^{k+1}} \right] \\ t_n &\approx \frac{(s_1 - 1)(1 - ps_1)}{(r + 1 - rs_1)q} \sum_{k=n+1}^{\infty} \frac{1}{s_1^{k+1}} \\ t_n &\approx \frac{(s_1 - 1)(1 - ps_1)}{(r + 1 - rs_1)q} \frac{1}{s_1^{n+2}} \frac{1}{1 - \frac{1}{s_1}} \\ t_n &\approx \frac{(s_1 - 1)(1 - ps_1)}{(r + 1 - rs_1)q} \frac{1}{s_1^{n+2}} \frac{s_1}{s_1 - 1} \\ t_n &\approx \frac{(1 - ps_1)}{(r + 1 - rs_1)q} \frac{1}{s_1^{n+1}} \end{aligned}$$

Problems

1. **Problem** Assume $N(s)$ and $D(s)$ in 1.8 have a common root s_1. Then write $N(s) = (s - s_1)N^*(s)$ and $D(s) = (s - s_1)D^*(s)$, where $N^*(s)$ and $D^*(s)$ are now polynomials with degree one less than $N(s)$ and $D(s)$, respectively. Use this presentation to show that derivation of 1.9 remains valid.

   Solution: $D^*(s)$ and $N^*(s)$ fulfill the assumptions under which 1.9 was proved. To show that 1.9 is still valid for $D(s)$ and $N(s)$ with common root s_1, we need to show that $\rho_1 = 0$ and ρ_i remains unchanged for $i = 2, 3, \ldots, m$.

   Noting that $D'(s) = D^*(s) + (s - s_1)D^{*\prime}(s)$, we see that $D'(s_1) \neq 0$, and because $N(s_1) = 0$, indeed $\rho_1 = 0$. Furthermore,

$$\frac{N(s_2)}{D'(s_2)} = \frac{(s_2 - s_1)N^*(s_2)}{D^*(s_2) + (s_2 - s_1)D^{*\prime}(s_2)} = \frac{N^*(s_2)}{D^{*\prime}(s_2)}.$$

   and similarly for $s_3, \ldots$, as was to be shown.

2. **Problem** Let $D(s) = a_0 + a_1 s + a_2 s^2 + \ldots a_{m-1} s^{m-1}$, $m \geq 1$, be a polynomial of degree $m - 1$ and assume $D(s)$ has m distinct roots. Show that then necessarily $D(s)$ is the null polynomial $D(s) = 0$ for all s. Hint: Use the identity

$$\begin{aligned}&D(t) - D(s) = \\ &\quad (t-s)[a_1 + a_2(t+s) + a_3(t^2 + ts + s^2) + \dots \\ &\quad + a_{m-1}(t^{m-2} + t^{m-3}s + \dots s^{m-2}]\end{aligned}$$

and prove the result by induction.

Solution:
When $m = 1$, $D(s) = a_0$ has one root, $\Rightarrow D(s) = 0$
Assume $m = k$ holds,
when $m = k+1$, $D(s)$ has $r_1, r_2, \dots, r_k + 1$ roots:

$$D(r_1) - D(s) = (r_1 - s)[a_1 + a_2(r_1 + s) + a_3(r_1^2 + r_1 s + s^2) + \dots + a_k(r_1^{k-1} + r_1^{k-2}s + \dots s^{k-1}]$$

Let $f(s) = [a_1 + a_2(r_1 + s) + a_3(r_1^2 + r_1 s + s^2) + \dots + a_k(r_1^{k-1} + r_1^{k-2}s + \dots s^{k-1}]$,
we can see that $f(s)$ is a polynomial of degree $k-1$ with k roots, so $f(s) = 0$

$$\Rightarrow D(r_1) - D(s) = 0$$

$$\Rightarrow D(s) = 0$$

3. **Problem** Sketch $D(s) = 1 - s + qp^r s^{r+1}$. Hint: Evaluate $D(s)$ at $s = -1/p$, $s = 1$, and $s = 1/p$ and distinguish the cases of r odd and r even.

   Solution: For r is odd, we should get a convex function with two roots. When r is even, we have three roots. In this case, when s is positive, we get a convex function. And if s is negative, the function is concave. You can sketch this function by identifying the roots and the limits of the function.

4. **Problem** Use your result of the previous problem to guide your numerical calculations for t_n as given in equation 1.20. In a commentary on possible over-reporting of leucine zipper motifs, Brendel & Karlin [1] used this method to derive benchmark values for the chance of observing a periodic leucine pattern in random sequences of typical protein lengths. Verify the numbers shown in the table of the commentary. Use sampling of computer-generated random sequences to empirically verify the mean and variance of waiting times given in 1.17 and 1.18 as well the approximation for the probability t_n for a few examples.