Chapter 2 – Discrete Distributions

We'll spend 6 classes on this 56-page chapter, and the results given herein for discrete distributions will be extended to continuous distributions in the next chapter (where sums will be replaced by integrals).

Discrete Probability Distributions (Section 2.1, pp. 40-47)

A box contains 10 chips: one [1], two [2]'s, three [3]'s and four [4]'s. If we choose one chip from this box, the *outcome space* is S = {[1], [2], [3], [4]}. Now define the *random variable* X which yields

- **the number 1 for the face [1]**
- **the number 2 for the face [2]**
- **the number 3 for the face [3]**
- **the number 4 for the face [4]**

Then, assuming the chips are each equally likely, we would assign the probability $\frac{1}{10}$ with the occurrence {X = 1}, the probability $\frac{2}{10}$ with the occurrence {X = 2}, etc. Next, define the probability mass function (denoted 'pmf') in general as f(x) = P(X = x). In this case, the pmf is:

$$f(x) = \frac{x}{10} \text{ for } x = 1, 2, 3 \text{ and } 4.$$

The graph of the *probability histogram* is on p.41 (dashed curve): here, f(1) = 0.10, f(2) = 0.20, etc. This is an example of a *discrete distribution* since the support of x is finite (at most countably infinite is okay); in the next chapter, we'll consider *continuous distributions*, which have an uncountably infinite support.

Quite different from the above theoretical pmf result, the text also reports the result of a sample of size n = 1000 from this population which results in the frequencies 98, 209, 305, 388 for x = 1, 2, 3, 4. These sample results give the relative frequency function h(1) = 0.098,

h(2) = 0.209, h(3) = 0.305, h(4) = 0.388 (very close to the pmf results given above) and graphed on p.41 via the darkened histogram. Note that in both cases, Σf(x) = Σh(x) = 1.

<u>Some Definitions</u>. The (theoretical) mean or expected value associated with the random variable X with pmf f(x) is μ = Σxf(x), where the sum is over all values of x. The sample mean is $\bar{x}$ = Σxh(x); another way to write this latter expression is:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{k=1}^{n} x_k}{n}$$

Whereas the mean is a good measure of center of a distribution, the spread of the distribution can be assessed by finding its variance (σ^2). The variance is obtained as follows: $\sigma^2 = \Sigma(x-\mu)^2 f(x)$, and the short-cut formula is $\sigma^2 = \Sigma x^2 f(x) - \mu^2$. Although the sample 'MLE' counterpart is $\nu = \Sigma x^2 h(x) - \bar{x}^2$,we will instead use the sample variance:

$$s^2 = \frac{n}{n-1}\nu = \frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n-1} = \frac{\left(\sum x_k^{\,2}\right) - \frac{1}{n}\left(\sum x_k\right)^2}{n-1}$$

Finally, the r^{th} moment about the origin is $\Sigma x^r f(x)$, so μ is the first moment about the origin, and $\Sigma x^2 f(x) = \sigma^2 + \mu^2$ is the second moment about the origin. Definition of the r^{th} sample moment about the origin is straightforward.

For the above example, the mean is μ = 1×0.1 + 2×0.2 + 3×0.3 + 4×0.4 = 3.0. The second moment about the origin is 1^2 ×0.1 + 2^2×0.2 + 3^2×0.3 + 4^2×0.4 = 10.0, so the variance is $\sigma^2 = 10.0 - 3.0^2$ = 1.0. On the other hand, the sample mean is $\bar{x}$ = 1×0.098 + 2×0.209 + 3×0.305 + 4×0.388 = 2.983. Also, at the bottom of p.43, the authors demonstrate that the sample variance is s^2 = 0.990.

Example 2.1-1 on p.44 is notable since it shows that the mean $\mu = 3.5$ does not need to be one of the support values of x. **Example 2.1-3** is relevant since for the pmf with X, $\sigma_X = \sqrt{2/3}$, whereas for the pmf of the more spread-out Y, $\sigma_Y = 2\sqrt{2/3}$ (twice as much as for X) – thus, σ does indeed capture the spread. **Example 2.1-5** on pp.45-7 illustrates the **Geometric distribution** (see p.64). In order to prove the pmf

$$f(x) = (1/3) \times (2/3)^{x-1},\ x = 1, 2, 3, \ldots$$

is a valid pmf and to find the mean and variance, we need to recall from Calculus that for $-1 < z < 1$,

$$1 + z + z^2 + z^3 + \ldots = \frac{1}{1-z}$$

$$1 + 2z + 3z^2 + 4z^3 + \ldots = \frac{1}{(1-z)^2}$$

$$1 + 2^2 z + 3^2 z^2 + 4^2 z^3 + \ldots = \frac{1+z}{(1-z)^3}$$

For Example 2.1-5, the mean is $\mu = 3$ and variance is $\sigma^2 = 6 = 2.45^2$, so the *expected number of rolls* of the die until the first [5] or [6] is **three (3)** give or take **2.45** rolls.

Expectations (Section 2.2, pp. 49-57)

As pointed out in the previous section, the (theoretical) mean of a distribution associated with the RV X is also called the **expected value** of X; it is denoted **E(X)** and μ. The Remark on p.50 shows that the mean μ does not always exist (since the series $1 + \frac{1}{2} + \frac{1}{3} + \ldots$ diverges); the authors also find the value of 'c' to make this a valid pmf.

We can find the *expected value of any real-valued function* **u(x)** by:

$$E[u(X)] = \Sigma u(x) f(x),$$

where the sum is again over all values of x.

Results (pp.52-53).

1. For the constant k, $E(k) = k$
2. $E[kv(X)] = kE[v(X)]$
3. $E[k_1v_1(X) + k_2v_2(X)] = k_1E[v_1(X)] + k_2E[v_2(X)]$
4. The previous results generalizes: $E[\Sigma k_m v_m(X)] = \Sigma k_m E[v_m(X)]$, and thus the expectation operator E is a *linear operator*.
5. For RVs X and Y with respective means μ_X and μ_Y and such that $Y = aX+b$, we have $\mu_Y = a\mu_X+b$ and $\sigma_Y^2 = a^2\sigma_X^2$.
6. On p.81, we'll show that if $X_1, X_2, \ldots X_n$ are independent RVs such that the mean and variance of X_k are μ_k and σ_k^2 for $k = 1\ldots n$ and $X = a_1X_1 + a_2X_2 + \ldots + a_nX_n$, then $\mu_X = a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$ and $\sigma_X^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2$.

Some applications are given on pp.52-53. Another application is to let the *independent* RVs $X_1, X_2 \ldots X_n$ have a Bernoulli distribution, which assigns the probability p to x = 1 and the probability q = 1 - p to x = 0. Thus, for each X_k, $\mu = 1\times p + 0\times(1-p) = p$; since $E(X^2) = 1^2\times p + 0^2\times(1-p) = p$, the variance is $\sigma^2 = p - p^2 = p(1-p)$.

Our interest here is in the RV $X = X_1+X_2+\ldots+X_n$, which counts the number of "successes" in the n independent Bernoulli trials; X has a Binomial distribution with parameters 'n' and 'p' – and we write $X \sim b(n,p)$. From the above Result 6, we know that $\mu_X = np$ and $\sigma_X^2 = np(1-p)$. Furthermore, the pmf of X is:

$$f(x) = \binom{n}{x} p^x(1-p)^{n-x} \text{ for } x = 0, 1, 2 \ldots n$$

On pp.54-55, the authors use results from the binomial expansion to show that this is a valid pmf (i.e., it sums to one), that $E(X) = \mu = np$, and that $E\{X(X-1)\} = n(n-1)p^2$. It follows that $E(X^2) = n(n-1)p^2 + np$, so $\sigma^2 = n(n-1)p^2 + np - n^2p^2 = np(1-p)$. [Note that the variance of this sum of n independent RVs is the sum of the n individual variances; more on that in Section 2.5.]

Students should become familiar with working with Table II on pp. 323-327. For example, if a coin is such that p = P(head) = 0.20 and we toss the coin n = 15 times, $\mu = 15\times0.20 = 3$ and $\sigma^2 = 15\times0.20\times0.80 = 2.4 = 1.55^2$, so we expect 3 heads give or take 1.55 heads. Also, the probability that we see *exactly 3 heads* is equal to 0.6482 – 0.3980 = 0.2502 = $\binom{15}{3}(0.2)^3(0.8)^{12}$ and the probability that we see *at least 3 heads* is 1 – 0.3980 = 0.6020.

Example 2.2-6 on p.56 raises the idea that in practice we wouldn't usually know the p in the binomial distribution and we need to estimate it (by using an *estimator*). If X ~ b(n,p), we let Y = $\frac{1}{n}$X, so that $\mu_Y = E(Y) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$ and $\sigma_Y^2 = \frac{1}{n^2}\sigma_X^2 = \frac{1}{n^2}[np(1-p)]$ so that $\sigma_Y^2 = \frac{p(1-p)}{n}$. Since E(Y) = p, we say that Y = $\frac{1}{n}$X is an unbiased estimator of p; more on this in Section 2.4.

Example 2.2-7 shows that the expression $E[(X-b)^2]$ is minimized when b is chosen to equal $\mu = E(X)$. Indeed, $E[(X-b)^2] = \sigma_X^2 + (\mu-b)^2$.

Chebyshev's Inequality states that for any positive constant k,

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

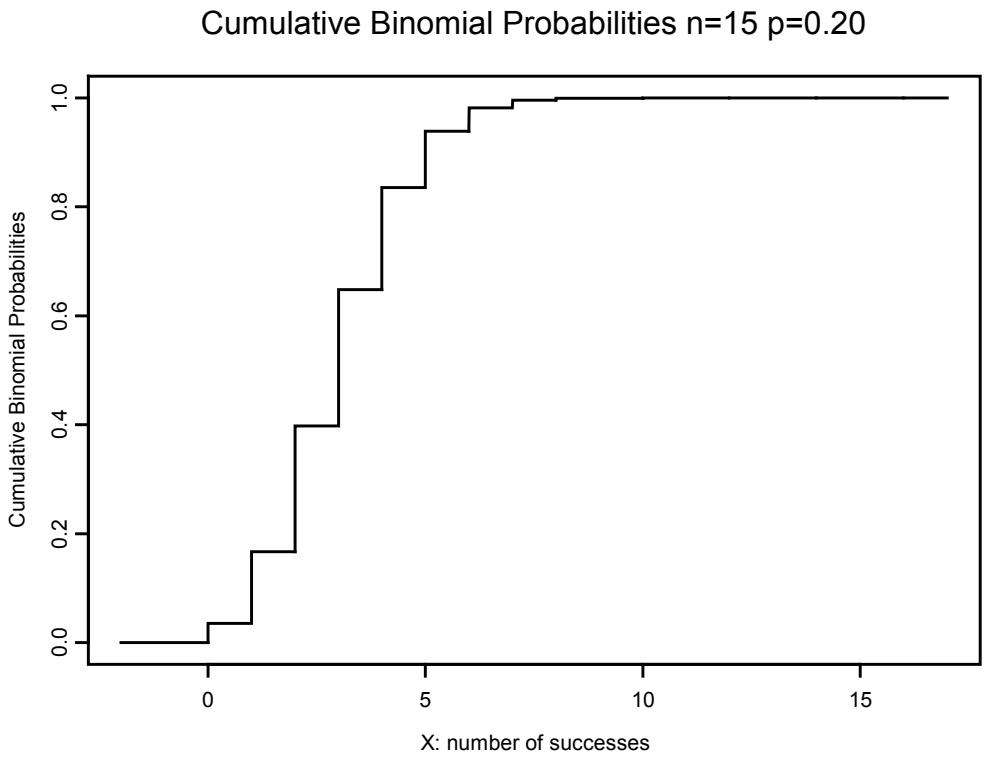
This applies to X = <u>any</u> random variable! Equivalently,

$$P(|X - \mu| < k\sigma) \geq 1 - \frac{1}{k^2}$$

For example (with k = 2), $P(\mu - 2\sigma < X < \mu + 2\sigma) \geq 0.75$.

On p.57, the authors use Chebyshev's Inequality to show that the limit of probabilities assessing the difference between Y = X/n and p is zero; we then say that Y = X/n converges in probability to p.

Special Discrete Distributions (Section 2.3, pp. 59-66)



In general, the cdf (cumulative density function) for X is defined as

$$F(x) = P(X \leq x)$$

For a discrete RV, all of the probability mass occurs at discrete points x_1, x_2 …, so the associated cdf is an increasing step function ranging from 0 to 1. Thus, by definition once F(x) reaches 1, it stays there. It's very important to note that *it is continuous from the right (but not necessarily from the left).*

On p.59, the cdf concept is illustrated for the Binomial distribution, and another example is shown above, here for X ~ b(15,0.20). The above graph is a step function and the change in the height from one step to the next is equal to $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$. Students are asked to work through Examples 2.3-1 – 2.3-3 and become very familiar with working with Table II on pp.323-7 (important for tests/quizzes).

Another illustration of a discrete RV distribution is the Poisson distribution which has the pmf:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, \text{ for } x = 0, 1, 2 \ldots$$

Note that this distribution has only one parameter (λ). Proof that this pmf sums to one uses the Maclaurin expansion, $e^z = \sum_{k=0}^{\infty} \frac{z^k}{k!}$, and this series also is used to show that for this distribution $\mu = \sigma^2 = \lambda$.

The Poisson distribution is a very important distribution in practice when we are counting the number of occurrences of some event in a fixed amount of time and where the probability of an event in a length of time is proportional to the interval length. Another application of the Poisson distribution is that it approximates the Binomial distribution for large n and small p (but with constant λ = np).

Example 2.3-6. X ~ poi(1) where X is number of flaws per 1200 ft. of computer tape. Then, from Table III on pp. 328-330, $P(X \le 2) = 0.920 = \frac{1^0 e^{-1}}{0!} + \frac{1^1 e^{-1}}{1!} + \frac{1^2 e^{-1}}{2!}$. If you count Y = the number of flaws occurring in 4800 ft. of tape (instead of 1200 ft.), it follows that Y ~ poi(4).

Example 2.3-5. X ~ b(100,0.05) and $P(X \le 1) = \binom{100}{0} 0.05^0\, 0.95^{100} + \binom{100}{1} 0.05^1\, 0.95^{99} =$ *0.0371*. Let's approximate using Y ~ poi(5). From Table III, we get 0.040; also, $\frac{5^0 e^{-5}}{0!} + \frac{5^1 e^{-5}}{1!} =$ *0.0404*, pretty close!

Let the RV X be the number of Bernoulli trials until the r^{th} success. Then X has the Negative Binomial distribution, and we can write X ~ nb(p); p here is the success probability. The pmf of X is:

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \text{ for } x = r, r+1, \ldots$$

The mean and variance are $\mu = \frac{r}{p}$ and $\sigma^2 = \frac{r(1-p)}{p^2}$. A special case is the Geometric distribution for which r = 1, where we are waiting for (counting the number of trials until) the first success.

Example 2.3-7. Success probability is p = 0.90 (90%), and the pmf for X = number of throws until the 10th successful free throw is at the bottom of p.64; g(10) = g(11) = 0.3487; and the expected number of shots until the 10th success is 11.11 give or take 1.111 $= \sqrt{1.235}$.

A box contains two types of objects, N_1 of the first type, N_2 of the second type, and so $N = N_1 + N_2$ total objects. We take a sample of size n from this box *without replacement* and let RV X = x be the number of items in the sample of the first type. The constraints on x are: $x \le n$, $x \le N_1$, and $n - x \le N_2$. Then the pmf of X is:

$$f(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$$

X has the Hypergeometric distribution, and the mean and variance are $\mu = np$ and $\sigma^2 = np(1-p)\frac{N-n}{N-1}$. In these expressions, $p = \frac{N_1}{N}$ is the proportion of the first type of objects in the box and $1-p = \frac{N_2}{N}$ is the proportion of the second type of objects in the box.

Example 2.3-8. N = 50 fuses of which $N_1 = 5$ are defective; we sample n = 4. The pmf is on the top of p.66 for x = 0, 1 … 4. The prob. of seeing no defective objects: f(0) = 0.647; prob. of at least one = 0.353.

Estimation (Section 2.4, pp. 68-74)

As is mentioned on p.56 in the context of a (Binomial) b(n,p) RV, $Y = \frac{1}{n}X$ is an *unbiased estimator* of p since E(Y) = p. How good is this estimator, and what makes an estimator "good"? What can we say about how close $Y = \frac{1}{n}X$ is to the parameter p?

Some Definitions. In the previous example, the parameter space is $\Omega = \{p: 0 < p < 1\}$. For the Poisson situation with pmf $f(x) = \frac{\theta^x e^{-\theta}}{x!}$, the parameter space is $\Omega = \{\theta: 0 < \theta < \infty\}$. A random sample of size n from this distribution means that the RVs $X_1, X_2 \ldots X_n$ are "iid" = independent and identically distributed, here each $X_k \sim$ Poisson(θ). A function of the RVs – called a statistic – is used to estimate θ – it's called an estimator, in this case a point estimator. When the realized values are used, we've calculated the point estimate. Due to assumed independence (iid), the joint pmf of $X_1, X_2 \ldots X_n$ is

$$P(X_1=x_1, X_2=x_2 \ldots X_n=x_n) = f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta)$$

Normally, we view this function as a function of $x_1, x_2 \ldots x_n$, but when the study is completed and the data is available, this can be viewed as a function of the unknown parameter θ. When we do so, we write L(θ), where L(*) denotes the likelihood function. Sometimes, it easier to just deal with LL(θ) = ln{L(θ)}, the log-likelihood function. The value of θ that maximizes L(θ) also maximizes LL(θ), and this value is called the maximum likelihood estimate (MLE) of θ.

Example 2.4-0. Suppose that $X_1, X_2 \ldots X_n$ are a random sample from a Bernoulli(p) distribution, so each $f(x;p) = p^x(1-p)^{1-x}$ for x=0 or x=1. Thus, the joint pmf is $f(x_1, x_2, \ldots, x_n)$ is obtained by evaluating this at each of the x's and multiplying; when this expression is viewed as a function of p, we obtain the likelihood function:

$$L(p) = p^{\Sigma x}(1-p)^{n-\Sigma x}$$

The sum here is over all values of x. The MLE can then be obtained by either maximizing L(p) or the log-likelihood expression:

$$LL(p) = (\Sigma x)\ln(p) + (n - \Sigma x)\ln(1-p),\ 0 < p < 1$$

Thus, the MLE here is $\hat{p} = \dfrac{\sum X_k}{n} = \overline{X}$

Example 2.4-1. Suppose that $X_1, X_2 \ldots X_n$ are iid Poisson(θ) RVs, so the log-likelihood function here is:

$$LL(\theta) = (\Sigma x)\ln(\theta) - n\theta - \ln(x_1!\, x_2! \ldots x_n!),\ \theta > 0$$

Thus, the MLE is $\hat{\theta} = \overline{X}$**, and that this is the maximum instead of just an extremum is confirmed by finding the second derivative and verifying that it evaluated at** $\hat{\theta} = \overline{X}$ **is negative.**

Example 2.4-2. Suppose that $X_1, X_2 \ldots X_n$ are a random sample from a Geometric(p) distribution, so the log-likelihood function here is:

$$LL(\theta) = n\ln(p) + (\Sigma x - n)\ln(1-p),\ 0 < p < 1$$

By differentiation, the MLE is $\hat{p} = \dfrac{n}{\sum X_k} = \dfrac{1}{\overline{X}}$**, which makes sense since we see n success in a total of Σx_k trials.**

Example 2.4-3. Suppose that $X_1, X_2 \ldots X_n$ are a random sample from a discrete uniform distribution over the integers 1, 2, … θ (where θ is unknown). Thus, the likelihood here is $L(\theta) = 1/\theta^n$ and so the log-likelihood function is $LL(\theta) = -n\ln(\theta)$ for θ > 0. By differentiation, $LL'(\theta) = -n/\theta$ which is < 0 for all values of θ. It follows that we must choose the smallest value possible for θ, so the MLE is

$$\hat{\theta} = \max\{X_1, X_2, \ldots, X_n\}$$

Example 2.4-4 (Capture-Recapture Problem). Estimating the total size of a population (e.g. of wolves in Yellowstone, fish in a lake) is an important problem in Estimation. We catch N_1 objects (wolves, fish, etc.) on the first instance, then tag and release them. After a given amount of time, we return to the venue and catch n objects of which x have tags on them from our first instance. One would then expect:

$$\frac{x}{n} = \frac{N_1}{N}$$

Our goal is to estimate N, the total number of objects (e.g., wolves in the park). Here, the relevant distribution is the Hypergeometric:

$$f(x) = \frac{\binom{N_1}{x}\binom{N-N_1}{n-x}}{\binom{N}{n}}$$

with the usual constraints on x, n, N_1 and N. Finding the MLE of N is very difficult due to the factorials. On p.74, it is shown to be the largest integer less than or equal to $\frac{N_1 n}{x}$. For example, if on the first instance we trap and tag $N_1 = 120$ wolves, and on the second we trap $n = 57$ wolves – $x = 22$ of which have tags – then since $N_1 n / x = 310.9$, our MLE is $\hat{N} = 310$ total wolves.

Application (pp. 75-78)

Suppose that the empirical results of a sample of n = 500 yields:

x	0	1	2	3	4
f	351	118	27	3	1

These data could have come from a Binomial(500,p), a Poisson(λ) or a translated Negative Binomial(r,p) distribution (or others). Here, we

will need to assume each of these distributions in turn and estimate the respective parameters. For this task, note that for these data, the sample statistics are $\bar{x} = 0.370$ and $s^2 = 0.402$.

In the Poisson distribution case, since the MLE of λ is the sample mean, we propose the Poisson distribution with $\lambda = 0.370$, calculate the probabilities P(0), P(1) … P(4), and multiply each of these by 500 to get the expected frequencies. These are given in the table at the top of p.77 and compared with the actual observed frequencies (not bad).

In the translated Negative Binomial case, equating the sample and theoretical means and variances, we get

$$0.370 = \frac{r(1-p)}{p} \qquad 0.402 = \frac{r(1-p)}{p^2}$$

Solving these (after rounding r to the nearest integer), we get the estimates r = 4 and p = 0.915. The pmf is then given on p.76 line 20, and is used to find P(0), P(1) … P(4), and the expected frequencies for this distribution (see the second table at the top of p.77); these latter estimated values are closer to the actual (observed values) than for the Poisson, so we would guess that the translated Negative Binomial distribution fits these data better.

Incidentally, the Binomial distribution is rejected for these data since for this distribution, we would then have (equating sample and theoretical means and variances)

$$0.370 = np \qquad 0.402 = np(1-p),$$

which implies that $1-p = 0.402/0.370 = 1.086$ and p = -0.086 (clearly impossible).

Linear Functions of Independent RVs (Section 2.5, pp. 79-83)

Returning iid RVs, the joint pmf of $X_1, X_2 \ldots X_n$ is

$$P(X_1=x_1, X_2=x_2 \ldots X_n=x_n) = f(x_1;\theta)\, f(x_2;\theta) \ldots f(x_n;\theta)$$

Maintaining the *independence* assumption but relaxing the *identically distributed* assumption, the joint pmf of $X_1, X_2 \ldots X_n$ is

$$P(X_1=x_1, X_2=x_2 \ldots X_n=x_n) = P(X_1=x_1)\,P(X_2=x_2) \ldots P(X_n=x_n)$$

$$= f_1(x_1;\theta)\,f_2(x_2;\theta) \ldots f_n(x_n;\theta)$$

For example, if $X_1 \sim$ poi($\lambda_1=2$); $X_2 \sim$ poi($\lambda_2=3$); X_1 and X_2 are independent, then as shown on p.80, $P(X_1+X_2=2) = \frac{5^2 e^{-5}}{2!}$. This suggests that the linear function $Y = X_1+X_2 \sim$ poi($\lambda_1+\lambda_2$); more on this in Chapter 6.

On pp.80-1, it is shown that for the linear combination $Y=a_1X_1+a_2X_2$ where X_1 and X_2 are independent, the mean of Y is $\mu_Y = a_1\mu_1+a_2\mu_2$ and variance of Y is $\sigma_Y^2 = a_1^2\sigma_1^2+a_2^2\sigma_2^2$. Also, it is shown that for independent RVs $X_1, X_2, \ldots X_n$ (with mean and variance of X_k equal to μ_k and σ_k^2 for k = 1…n), then the mean of $Y = a_1X_1+a_2X_2+\ldots+a_nX_n$ is $\mu_Y=a_1\mu_1+a_2\mu_2+\ldots+a_n\mu_n$; its variance is $\sigma_Y^2=a_1^2\sigma_1^2+a_2^2\sigma_2^2+\ldots+a_n^2\sigma_n^2$

In the text's proof of the above on p.80, it is also pointed out that if X_1 and X_2 are independent, then (→ but not necessarily ←):

(1) $E[(X_1-\mu_1)(X_2-\mu_2)] = E(X_1-\mu_1)\,E(X_2-\mu_2) = 0$
(2) $E[u_1(X_1)\,u_2(X_2)] = E[u_1(X_1)]\,E[u_2(X_2)]$

To illustrate, X_1 and X_2 are independent, $X_1 \sim b(100,½)$ and $X_2 \sim b(48,¼)$, and $Y = X_1-X_2$, then the mean is $\mu_Y = 100\times½ - 48\times¼ = 38$ and the variance is $(1)^2\times100\times½\times½ + (-1)^2\times48\times¼\times¾ = 34$. Also, if $W = X_1X_2$, then $E(W) = 50\times12 = 600$.

Returning to iid RVs $X_1, X_2 \ldots X_n$ each with mean μ and variance σ^2, then the RV $\overline{X} = \frac{1}{n}(X_1+X_2+\ldots+X_n)$ has expected value $\mu_{\overline{X}} = \mu$ and

variance $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$. On p.82, the authors show that $\bar{X}$ converges in probability to μ, illustrating the WLLN (weak law of large numbers); on p.57, it was also shown that for X ~b(n,p), $\frac{X}{n}$ converges in probability to p. Thus, in some sense, the unbiased estimators $\bar{X}$ and $\frac{X}{n}$ are good estimators for μ and p respectively.

The standard error associated with $\bar{X}$ is $s_{\bar{X}} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$ (called the SEM or standard error of the mean), and so an approximate (so-called Wald) 95% confidence interval (CI) for μ is thus $\bar{x} \pm 2\frac{s}{\sqrt{n}}$.

With regard to estimating a Binomial proportion (see p.56), the standard error (SEP) associated with $\frac{X}{n}$ is $s_{X/n} = \sqrt{\frac{(X/n)(1-X/n)}{n}}$, so the approximate 95% Wald CI for p is $\frac{x}{n} \pm 2\sqrt{\frac{(x/n)(1-x/n)}{n}}$. Illustrations are given in Examples 2.5-4 and 2.5-5 on p.83; note that in Ex. 2.5-4, the CI for p contains ½ = 50% – what is the relevance of this?

Multivariate Discrete Distributions (Section 2.6, pp. 84-93)

In this section, we'll relax the above independence assumption as well. In this case, the full joint pmf needs to be written out in functional or tabular form. An example appears on p.85:

		x_1		
		1	2	$f_2(x_2)$
x_2	2	3/10	1/10	4/10
	1	4/10	2/10	6/10
$f_1(x_1)$		7/10	3/10	1

The (black) probabilities in the center of the table are the joint probabilities; for example, f(1,2) = 3/10. The (red and blue) probabilities in the margins are the marginal pmf's: for example, $f_2(1) = 0.60$ and $f_2(2) = 0.40$. Note that it is not true that for all x_1 and x_2, $f(x_1,x_2) = f_1(x_1)f_2(x_2)$; thus X_1 and X_2 are dependent.

For $Y = u(X_1,X_2)$, $E(Y) = E[u(X_1,X_2)] = \sum_{x_1}\sum_{x_2} u(x_1,x_2)f(x_1,x_2)$; it follows that $E[u(X_1)] = \sum_{x_1} u(x_1)f_1(x_1)$ and $E[u(X_2)] = \sum_{x_2} u(x_2)f_2(x_2)$ where $f_1(x_1) = \Sigma f(x_1,x_2)$ (summing only over x_2) and $f_2(x_2) = \Sigma f(x_1,x_2)$ (summing only over x_1). Thus, both means and variances can be found for the marginal distributions. For the above example, $\mu_1 = E(X_1) = (1)(7/10) + (2)(3/10) = 1.3$ and $\sigma_1^2 = Var(X_1) = 0.21 = 0.4583^2$. Similarly, $\mu_2 = 1.4$, $\sigma_2^2 = 0.24 = 0.4899^2$; also, $E(X_1X_2) = (1)(1)(0.40) + (1)(2)(0.30) + (2)(1)(0.20) + (2)(2)(0.10) = 1.80$.

Important conditions related to the joint pmf are given on p.86. Also, the covariance of X_1 and X_2 is $\sigma_{12} = Cov(X_1,X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$, and it's easy to show that $\sigma_{12} = E(X_1X_2) - \mu_1\mu_2$; the correlation coefficient of X_1 and X_2 is $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$. In the last example, $\sigma_{12} = 1.80 - (1.3)(1.4) = -0.02$ and $\rho = -0.02/(0.4583*0.4899) = -0.0891$. Students should also work through Example 2.6-3 on p.87.

If we have a sample of 2-ples of size n, (x_1,y_1), (x_2,y_2), … (x_n,y_n), then the sample correlation coefficient (r) is

$$r = \frac{\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^{n}(x_k - \bar{x})^2}\sqrt{\sum_{k=1}^{n}(y_k - \bar{y})^2}}$$

A short-cut formula for r is given at the top of p.88. Extending the above results to n possibly correlated RVs $X_2, \ldots X_n$ (with respective means and variances μ_k and σ_k^2 for k = 1…n), then it's not hard to show that for $Y = \Sigma a_k X_k$, the mean is $\mu_Y = \Sigma a_k \mu_k$; the variance of Y is:

$$\sigma_Y^2 = \sum_{k=1}^{n} a_k^2 \sigma_k^2 + 2\sum_{k=1}^{n}\sum_{m=1}^{k-1} a_k a_m \sigma_{km}$$

See Example 2.6-4 on p.88.

The conditional pmf of X_1 given that $X_2 = x_2$ is defined by $g(x_1|x_2)$ (provided $f_2(x_2) > 0$):

$$g(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$

Similarly, the conditional pmf of X_2 given that $X_1 = x_1$ is

$$h(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

(provided $f_1(x_1) > 0$). Example 2.6-5 on p.89 provides an illustration: Here, $f(x_1,x_2) = \frac{1}{21}(x_1+x_2)$, $x_1 = 1,2,3$ and $x_2 = 1,2$; the marginal distributions $f_1(x_1)$ and $f_2(x_2)$ as well as the conditional distributions $g(x_1|x_2)$ and $h(x_2|x_1)$ are given on pp.89-90. For this joint density, the following table is relevant:

		x_1			
		1	2	3	$f_2(x_2)$
x_2	2	3/21 [3/12] (3/5)	4/21 [4/12] (4/7)	5/21 [5/12] (5/9)	12/21
	1	2/21 [2/9] (2/5)	3/21 [3/9] (3/7)	4/21 [4/9] (4/9)	9/21
$f_1(x_1)$		5/21	7/21	9/21	1

In the above table, the black probabilities are the joint probabilities, the red and blue probabilities are again the marginal probabilities, the green probabilities are the conditional probabilities given x_2, and the orange probabilities are the conditional probabilities given x_1.

Conditional *means* and *variances* are defined on p.91 in the logical manner. In Example 2.6-6 on p.91, note that $E[X_2|X_1=3] = (1)(4/9) + (2)(5/9) = 14/9$, $E[X_2^2|X_1=3] = (1^2)(4/9) + (2^2)(5/9) = 24/9$, whence $Var[X_2|X_1=3] = 20/81$. Note that these results differ from those for the marginal distribution for X_2 (since these variables are correlated): $E[X_2] = 11/7$ and $Var[X_2] = 12/49$.

Finally, noting that $E[X_2|X_1=x_1]$ is a function of x_1 alone, in the case where it is linear function of x_1, then we can write

$$E[X_2|X_1=x_1] = a + bx_1$$

On p.92, it is shown that for this linear case, the intercept and slope are $a = (\mu_2 - \rho\mu_1\frac{\sigma_2}{\sigma_1})$ and $b = \rho\frac{\sigma_2}{\sigma_1}$, so we can write:

$$E[X_2|X_1=x_1] = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1)$$

We return to these latter ideas when we study linear regression (Section 4.6).

It is pointed out on p.93, that similar results to those given in Section 2.6 apply for the continuous case using *integrals* instead of sums.

Additional comments appear on p.96. First, the Poisson Process postulates are given and discussed. Also, the contributions of Sir Ronald A. Fisher and Sir David Cox to Statistics are highlighted.