Chapter 2 - Questions and Solutions

Edwin Fennell

2.1 Derive the mean and variance for the binomial distribution

The binomial distribution $B(n,p)$ is defined on $\mathbb{N}$ by the probability mass function

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, ..., n \\ 0 & else \end{cases}$$

More abstractly, it is the distribution of total successes of n independent Bernoulli trials with success probability p. This immediately gives the expectation as

$$np$$

since the expectation of a sum of random variables is just the sum of the expectation. Since the Bernoulli trials are independent, the variance of their sum is also just the sum of their variances, so

$$np(1-p)$$

2.2 Derive the mean and variance for the uniform distribution

The uniform distribution $N(a,b)$ (for $a < b$) is defined on $\mathbb{R}$ by the probability density function

$$p(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & else \end{cases}$$

We therefore calculate the expectation as

$$\int_a^b \frac{x}{b-a} dx = \left[\frac{x^2}{2(b-a)} \right]_a^b = \frac{a+b}{2}$$

This result is also reasonably clear from a symmetry argument.

We also have

$$\int_a^b \frac{x^3}{b-a} dx = \left[\frac{x^3}{3(b-a)} \right]_a^b = \frac{a^2+ab+b^2}{3}$$

which gives us the variance as

$$\frac{a^2+ab+b^2}{3}-\frac{(a+b)^2}{4}=\frac{a^2-2ab+b^2}{12}=\frac{(b-a)^2}{12}$$

2.3 Derive the mean and covariance matrix for the multivariate normal distribution

The multivariate Gaussian $N(\mu,\Sigma)$ is defined on $\mathbb{R}^k$ by the probability density function

$$p(x)=\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

Then the ith element of the expectation is

$$\int x_i\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx$$

Σ is symmetric and positive definite, so by Cholesky composition we can write

$$\Sigma^{-1}=H^TH$$

If we set $y=H(x-\mu)$ then $dx=|H^{-1}|dy$, and changing variables in the above integral gives us

$$\int(H^{-1}_{ik}y_k+\mu_i)\frac{1}{(2\pi)^{\frac{k}{2}}}exp\left(-\frac{1}{2}y\cdot y\right)dy$$

We now use symmetry to discard the Hy and consider all elements to obtain our expectation as

$$\mu$$

The (i,j)th element of the covariance matrix is given by

$$\int(x_i-\mu_i)(x_j-\mu_j)\frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}}exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx$$

Performing the previous substitution gives

$$\int H^{-1}_{ia}y_aH^{-1}_{jb}y_b\frac{1}{(2\pi)^{\frac{k}{2}}}exp\left(-\frac{1}{2}y\cdot y\right)dy$$

This is where the real power of this substitution comes in - we could have derived the mean without it just using the symmetry around μ, but here we would have been stuck. Observe that the substitution has made our integral separable. For any fixed $(i.j)$ we write this integral as the sum of integrals indexed by a and b - note that the value of any constituent integral with $a! = b$ is necessarily 0 by symmetry in our separated integrals. Therefore we just sum over all cases where $a = b$, and play with indices in the result, to get the value of the integral as

$$\int (H^{-1}(H^T)^{-1})_{ij}(y \cdot y) \frac{1}{(2\pi)^{\frac{k}{2}}} exp\left(-\frac{1}{2} y \cdot y\right) dy$$

Using results for variance of a 1D-Gaussian and separating the integrand, this simplifies to

$$(HH^T)^{-1}_{ij} = \Sigma_{ij}$$

distribution And so the covariance matrix is just

$$\Sigma$$

2.4 Show that the mean and variance of the beta distribution with parameters a and b are

$$\boldsymbol{\frac{a}{a+b}}$$

and

$$\boldsymbol{\frac{ab}{(a+b)^2(a+b+1)}}$$

respectively

The beta distribution $Beta(a, b)$ is defined on $[0, 1]$ by the probability density function

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a, b)}$$

where

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

The expectation is given by

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

We note that $\Gamma(a+b+1) = (a+b)\Gamma$ and $\Gamma(a+1) = a\Gamma(a)$ so our integral becomes

$$\frac{a}{a+b} \int_0^1 \frac{\Gamma(a+b+1)x^a(1-x)^{b-1}}{\Gamma(a+1)\Gamma(b)} dx$$

The integrand is just the density function of $Beta(a+1, b)$, so the above expression just reduces to

$$\frac{a}{a+b}$$

Similarly, we have

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

We have $\Gamma(a+b+2) = (a+b+1)(a+b)\Gamma(a+b)$, $\Gamma(a+2) = (a+1)(a)\Gamma(a)$, so this integral is just equal to

$$\frac{a(a+1)}{(a+b)(a+b+1)} \int_0^1 \frac{\Gamma(a+b+2)x^{a+1}(1-x)^{b-1}}{\Gamma(a+2)\Gamma(b)}dx = \frac{a(a+1)}{(a+b)(a+b+1)}$$

and therefore the variance is

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

2.5 Show that the normalising constant in the beta distribution with parameters a,b is given by

$$\boldsymbol{\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}$$

The beta distribution $Beta(a,b)$ is defined on $[0,1]$ by the probability density function

$$\frac{x^{a-1}(1-x)^{b-1}}{B(a,b)}$$

We are required to prove that

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

This is equivalent to showing

$$\int_{\mathbb{R}} x^{a-1}(1-x)^{b-1}dx = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

since any probability distribution must integrate to 1 over its domain. The gamma function is defined by

$$\Gamma(a) = \int_{\mathbb{R}} x^{a-1}e^{-x}$$

so we can multiply through by $\Gamma(a+b)$ in the above expression and expand to obtain our proposition as an equality between two double integrals:

$$\int_{\mathbb{R}}\int_{\mathbb{R}} x^{a-1}(1-x)^{b-1}y^{a+b-1}e^{-y}dxdy = \int_{\mathbb{R}}\int_{\mathbb{R}} u^{a-1}e^{-u}v^{b-1}e^{-v}dudv$$

On the right-hand side we can substitute $u = xy$ and $v = (1-x)y$, which gives us the following relation between area elements:

$$dudv = \begin{vmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{vmatrix} dxdy = \begin{vmatrix} y & x \\ -y & (1-x) \end{vmatrix} dxdy = y \cdot dxdy$$

Using this and fully substituting leaves the right-hand side as

$$\int_{\mathbb{R}} \int_{\mathbb{R}} (xy)^{a-1}((1-x)y)^{b-1} y \cdot dxdy$$

which is just the left-hand side of the equation. $\square$

2.6 Show that the mean and variance of the pdf of the gamma distribution with parameters a,b

$$\boldsymbol{p(x) = \begin{cases} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} & x > 0 \\ 0 & else \end{cases}}$$

are given by

$$\boldsymbol{\frac{a}{b}}$$

and

$$\boldsymbol{\frac{a}{b^2}}$$

respectively.

The expectation is given by

$$\int_0^\infty \frac{b^a}{\Gamma(a)} x^a e^{-bx} dx$$

Setting $u = bx$ and substituting into the above integral yields

$$\int_0^\infty \frac{u^a}{b\Gamma(a)} e^{-u} du = \frac{\Gamma(a+1)}{b\Gamma(a)} = \frac{a}{b}$$

where the last equality comes from the relation $\Gamma(a+1) = a\Gamma(a)$

The variance is given by

$$\int_0^\infty \frac{b^a}{\Gamma(a)} x^{a+1} e^{-bx} dx - \frac{a^2}{b^2}$$

We make the same integral substitution as above which causes the expression to reduce to

$$\int_0^\infty \frac{u^{a+1}}{b^2\Gamma(a)} e^{-u} du - \frac{a^2}{b^2} = \frac{\Gamma(a+2)}{\Gamma(a)b^2} - \frac{a^2}{b^2} = \frac{a}{b^2}$$

since $\Gamma(a+2) = a(a+1)\Gamma(a)$

2.7 Show that the mean and covariance of a Dirichlet pdf with parameters $\boldsymbol{a_k}$, $\boldsymbol{k = 1, 2, ..., K}$ are given by

$$\boldsymbol{\frac{a}{\bar{a}}}$$

and

$$\boldsymbol{\frac{\bar{a}\delta_3 a - aa^T}{\bar{a}^2(1-\bar{a})}}$$

respectively

($\boldsymbol{\bar{a}}$ is the sum of all elements of $\boldsymbol{a}$ and $\boldsymbol{\delta_3}$ is the 3D Kronecker delta)

The pdf of this Dirichlet distribution over $\mathbb{R}^k$ is defined by

$$\begin{cases} \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)} \prod_{k=1}^K x_k^{a_k - 1} & (\forall i, 0 \leq x_i \leq 1) \wedge (\sum_{i=1}^K x_i = 1) \\ 0 & else \end{cases}$$

We won't write out complete integrals here because they are horrible - we will instead use some neat trickery to rewrite the integrands of expectations we want in terms of other Dirichlet pdfs, which conveniently integrate to 1 across our domain. We will therefore write our compound integral using a single integral sign over our domain D, and just write dx for our hypervolume element.

The ith component of the mean is given by

$$\int_D x_i \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)} \prod_{k=1}^K x_k^{a_k - 1} dx$$

If we define a vector b s.t. $b_j = a_j$ for $j \neq i$, and $b_i = a_i + 1$, then we note that the above integral is just

$$\int_D \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)} \prod_{k=1}^K x_k^{b_k - 1} dx$$

$\bar{b} = \bar{a} + 1$, so we can use the our nifty relation on gammas

$$\Gamma(x+1) = x\Gamma(x)$$

to simplify the above to

$$\frac{a_i}{\bar{a}} \int_D \frac{\Gamma(\bar{b})}{\Gamma(b_1)...\Gamma(b_k)} \prod_{k=1}^K x_k^{b_k - 1} dx$$

The integrand is just the pdf of the Dirichlet distribution with parameter vector b, and so the integral evaluates to 1. This gives our mean vector as just

$$\frac{a}{\bar{a}}$$

The i,jth element of the covariance is given by

$$\int_D x_i x_j \frac{\Gamma(\bar{a})}{\Gamma(a_1)...\Gamma(a_k)} \prod_{k=1}^{K} x_k^{a_k-1} dx - \frac{a_i a_j}{\bar{a}^2}$$

If $i \neq j$, we can define a vector c s.t. $c_k = a_k$ for $k \neq i,j$, and $c_k = a_k + 1$ for $k = i,j$. Then we again use our gamma relation to rearrange the above integral to

$$\frac{a_i a_j}{\bar{a}(\bar{a}+1)} \int_D \frac{\Gamma(\bar{c})}{\Gamma(c_1)...\Gamma(c_k)} \prod_{k=1}^{K} x_k^{c_k-1} dx = \frac{a_i a_j}{\bar{a}(\bar{a}+1)}$$

since the integrand above is just the pdf of the Dirichlet distribution with parameter vector c. This gives the value of the i,jth element as

$$\frac{a_i a_j}{\bar{a}} \left(\frac{1}{\bar{a}+1} - \frac{1}{\bar{a}} \right) = -\frac{a_i a_j}{\bar{a}^2(\bar{a}+1)}$$

If instead $i = j$, then define d s.t. $d_k = a_k$ for $k \neq i$ and $d_i = a_i + 2$. So our expression for the i,ith element becomes

$$\frac{a_i(a_i+1)}{\bar{a}(\bar{a}+1)} \int_D \frac{\Gamma(\bar{d})}{\Gamma(d_1)...\Gamma(d_k)} \prod_{k=1}^{K} x_k^{d_k-1} dx - \frac{a_i^2}{\bar{a}^2}$$

As with the previous examples, the integral just evaluates to 1, so this simplifies to

$$\frac{a_i}{\bar{a}(\bar{a}+1)} + \frac{a_i^2}{\bar{a}} \left(\frac{1}{\bar{a}+1} - \frac{1}{\bar{a}} \right) = \frac{\bar{a}a_i - a_i^2}{\bar{a}^2(\bar{a}+1)}$$

When we consider this across all i,j, we get the proposed expression for covariance.

2.8 Show that the sample mean of N i.i.d. samples, is an unbiased estimator with variance that tends to zero asymptotically as $N \to \infty$

Suppose we are sampling from an underlying distribution A, and our ith samples is denoted by A_i - all independent and distributed as A. Our sample mean is defined by

$$\bar{A} = \frac{1}{n} \sum_{i=1}^{K} A_i$$

Expectation distributes over all sums, therefore

$$E(\bar{A}) = E\left(\frac{1}{n}\sum_{i=1}^{K} A_i\right) = \frac{1}{n}\sum_{i=1}^{K} E(A_i) = \frac{n * E(A)}{n} = E(A)$$

since all the A_i are distributed as A is. Therefore the sample mean is unbiased. This tells us that the variance of the sample mean is given by

$$E((\bar{A} - E(A)(\bar{A} - E(A)))$$

Using again the distribution of expectation over addition and scalar multiples, we can rewrite this as

$$\sum_{i=1}^{N}\sum_{j=1}^{N} E\left(\frac{A_i - E(A)}{n}\frac{A_j - E(A)}{n}\right)$$

By independence of the A_i and the fact that all the A_i have mean $E(A)$, the cross-terms cancel and we are left with

$$\sum_{i=1}^{j} \frac{Var(A_i)}{n^2} = \frac{n * Var(A_i)}{n^2} = \frac{Var(A)}{n}$$

This converges to 0 as $N \to \infty$

2.9 Show that for WSS processes

$$\boldsymbol{r(0) \geq |r(k)|, \forall k \in \mathbb{Z}}$$

and that for jointly-WSS processes

$$\boldsymbol{r_u(0) r_v(0) \geq |r_{uv}(k)|, \forall k \in \mathbb{Z}[sic]}$$

A WSS (Wide-Sense Stationary) process is defined as one for which $\mu_n = \mu$ $\forall n$ for some μ and for which $r(n, n-k) = r(k)$ $\forall n, k$ ($r(a, b)$ is the autocorrelation at times a, b).

We define a process by the time-indexed sequence of random variables u_t, where t is time. Suppose that it is WSS. Consider the random variable $(u_k - u_0)^2$ for arbitrary k. It is non-negative, so must have non-negative mean. Therefore

$$0 \leq E((u_k - u_0)^2) = E(u_k u_k) - 2E(u_k u_0) + E(u_0 u_0) = 2r(0) - 2r(k)$$

This immediately gives $r(0) \geq r(k)$.

The second part of the assertion is not true (consider two processes with a constant value of $\frac{1}{2}$). The mathematical expression should probably read

$$r_u(0)r_v(0) \geq |r_{uv}(k)|^2$$

This follows from a special case of Hölder's inequality, which states that the product of the expectation of two non-negative random variables exceeds the square of the expectation of the product of those variables. We can use this elementwise on the absolulute values elements of u_0 and v_k and sum the resulting inequalities to obtain

$$E(u_0 \cdot u_0)E(v_k \cdot v_k) \geq (E(abs(u_0) \cdot abs(v_k))^2$$

where here $abs(a)$ is the vector containing the elementwise absolute values of a. $|abs(u_0) \cdot abs(v_k)| > |u_0 \cdot v_k|$ always, so

$$E(abs(u_0) \cdot abs(v_k)) = |E(abs(u_0) \cdot abs(v_k))| \geq |E(u_0 \cdot v_k)|$$

and the desired result follows immediately from this.

2.10 Show that the autocorrelation of the output of a linear system, with impulse response $w_n, n \in \mathbb{Z}_n$ is related to the autocorrelation of the input WSS process, via

$$\boldsymbol{r_d(k) = r_u(k) * w_k * w^*_{-k}[sic]}$$

This is wrong - you can't convolve over scalars. What is actually required to be shown is

$$r_d(k) = (r_u * w * \bar{w}^*)_k$$

where $\bar{w}$ is the sequence obtained by flipping w about the origin and r_u is a sequence defined by $(r_u)_i = r_u(i)$.

By definition

$$d_k = \sum_{i=-\infty}^{\infty} u_i^* w_{k-i}$$

Therefore we may write

$$r_d(k) = E(d_0 d_k) = E\left(\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} u_i^* w_{-i} u_j^* w_{k-j}\right)$$

Since j is in the inner loop, we can make the change of variables $a = j - i$ and obtain

$$E\left(\sum_{i=-\infty}^{\infty} \sum_{a=-\infty}^{\infty} u_i^* w_{-i} u_{a+i}^* w_{k-a-i}\right)$$

We take the expectation inside to obtain

$$\sum_{i=-\infty}^{\infty} \sum_{a=-\infty}^{\infty} E(u_i u_{a+i})^* w_{-i} w_{k-a-i} = \sum_{i=-\infty}^{\infty} \sum_{a=-\infty}^{\infty} r_u(a)^* w_{-i} w_{k-a-i}$$

We can now switch the sums, use the fact that $r_u(a)$ is independent of i, and subtract k from the indices of the inner sum by k (which doesn't affect the summation indices) to obtain

$$\sum_{a=-\infty}^{\infty} r_u(a)^* \sum_{i=-\infty}^{\infty} w_{-i} w_{k-a-i}$$

We now denote $\bar{w}$ as the sequence w reversed about the origin, giving our expression as

$$\sum_{a=-\infty}^{\infty} r_u(a)^* \sum_{i=-\infty}^{\infty} \bar{w}_i w_{k-a-i}$$

This is just

$$\sum_{a=-\infty}^{\infty} r_u(a)^* (\bar{w}^* * w)_{k-a}$$

which is just

$$(r_u * w * \bar{w}^*)_k$$

as required.

2.11 Show that

$$\boldsymbol{ln(x) \leq x - 1}$$

Consider the function

$$f(x) = x - 1 - ln(x)$$

The statement we are required to prove is equivalent to f being everywhere non-negative. f is differentiable everywhere it is defined (on $(0, \infty)$), and so we have

$$f'(x) = 1 - \frac{1}{x}$$

We note that f' is negative for $x < 1$ and positive for $x > 1$. Therefore the minimum of f occurs at $x = 1$. $f(1) = 0$, therefore $f \geq 0$ everywhere. $\square$

2.12 Show that

$$\boldsymbol{I(X, Y) \geq 0}$$

The average mutual information $I(X,Y)$ is defined by

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

. We can restrict this sum to combinations of x, y s.t. $P(x,y) > 0$. This condition implies $P(x)P(y) > 0$ Now we can use the solution to problem 2.13, since

$$\sum_{x,y\ s.t.\ P(x,y)>0} P(x,y) = 1$$

and

$$\sum_{x,y\ s.t.\ P(x,y)>0} P(x)P(y) \leq 1$$

and $P(x,y) \geq 0, P(x)P(y) \geq 0\ \forall x, y$ we have

$$\sum_{x,y\ s.t.\ P(x,y)>0} P(x,y)log(P(x,y)) \geq \sum_{x,y\ s.t.\ P(x,y)>0} P(x,y)log(P(x)P(y))$$

which immediately rearranges to the required result.

2.13 Show that if $a_i, b_i, i = 1, 2, ..., M$ are positive numbers with

$$\boldsymbol{\sum_{i=1}^{M} a_i = 1, \sum_{i=1}^{M} b_i \leq 1}$$

then

$$\boldsymbol{-\sum_{i=1}^{M} a_i \log(a_i) \leq -\sum_{i=1}^{M} a_i \log(b_i)}$$

This follows straightforwardly from the concavity of log. Since the a_i are positive and sum to 1 we have

$$\sum_{i=1}^{M} a_i \log\left(\frac{b_i}{a_i}\right) \leq \log\left(\sum_{i=1}^{M} a_i \frac{b_i}{a_i}\right) = \log\left(\sum_{i=1}^{M} b_i\right) \leq \log(1) = 0$$

The initial proposition follows from a simple rearrangement.

2.14 Show that the maximum value of the entropy of a random variable occurs if all possible outcomes are equiprobable

Suppose that X is the random variable on M outcomes with the highest entropy (exists via compactness argument). Suppose X takes outcome x_i with probability p_i for $i = 1, 2, ..., M$. Now suppose that not all the p_i are equal. By reordering, $p_1 < p_2$. Now we define a random variable Y

taking the same values as X with the same probabilities, apart from x_1, x_2, which Y takes with probabilities $\frac{p_1+p_2}{2}$ each. We note that the function $f(x) = x\log(x)$ is strictly convex in the range $[0, 1]$ (second derivative is $1/x$) so we get

$$\frac{p_1 + p_2}{2} \log\left(\frac{p_1 + p_2}{2}\right) < \frac{1}{2}(p_1 \log(p_1) + p_2 \log(p_2))$$

which when we consider the expressions for entropy of both X and Y, tells us that X has a lower entropy than Y. This is a contradiction, so all the p_i must be equal. $\square$

2.15 Show that of all the p.d.f.s that describe a random variable in an interval $[a, b]$, the uniform one maximises the entropy.

We can use Jensen's inequality here - given a convex function ϕ and a real-valued random variable X, we have

$$\phi(E(X)) \leq E(\phi(X))$$

We note that $f(x) = x\log(x)$ is a convex function, and so given any p.d.f ψ, we have

$$E(\psi)\log(E(\psi)) \leq E(\psi\log(\psi))$$

Multiplying through by -1, and noting that $E(\psi) = \frac{1}{b-a}$, we directly obtain

$$entropy((b-a)^{-1}) \geq entropy(\psi)$$

ψ was arbitrary, so we have that the uniform distribution on $[a, b]$ has the greatest entropy of all distributions on $[a, b]$. $\square$