Spring 2013

1. Consider an independent identically distributed sequence $X_1, X_2, \ldots, X_{n+1}$ taking values 0 or 1 with probability distribution

$$P(X_i = 1) = 1 - P(X_i = 0) = p.$$

Uniformly choose M *fragments* $F_1, F_2, \ldots, F_M$ of length 2 starting in the interval $[1, n]$, that is, $F_i = (X_{j_1}, X_{j_1+1})$ for some $1 \leq j_i \leq n$. Let $\boldsymbol{W} = (1, 1)$.

a) Let $N_{\boldsymbol{W}}$ be the number of times the word $\boldsymbol{W}$ occurs among the M fragments. Calculate $\mathbb{E}(N_{\boldsymbol{W}})$.
Solution. We have

$$\mathbb{E}(N_{\boldsymbol{W}}) = \sum_{i=1}^{M} \mathbf{1}(X_{j_i} = 1, X_{j_i+1} = 1) = \boxed{Mp^2}.$$

b) Calculate the probability $P(F_1 = \boldsymbol{W}, F_2 = \boldsymbol{W})$.
Solution. First, we note that there are three cases:

i. $|\{X_{j_1}, X_{j_1+1}\} \cap \{X_{j_2}, X_{j_2+1}\}| = 2$ happens n times with $P(F_1 = \boldsymbol{W}, F_2 = \boldsymbol{W}) = p^2$.
ii. $|\{X_{j_1}, X_{j_1+1}\} \cap \{X_{j_2}, X_{j_2+1}\}| = 1$ happens $2n - 4$ times with $P(F_1 = \boldsymbol{W}, F_2 = \boldsymbol{W}) = p^3$.
iii. $|\{X_{j_1}, X_{j_1+1}\} \cap \{X_{j_2}, X_{j_2+1}\}| = 0$ happens $n^2 - 3n + 4$ times with $P(F_1 = \boldsymbol{W}, F_2 = \boldsymbol{W}) = p^4$.

Hence, combining the three cases, we get

$$P(F_1 = \boldsymbol{W}, F_2 = \boldsymbol{W}) = \boxed{\frac{p^2(n + (2n-4)p + (n^2 - 3n + 4)p^2)}{n^2}}.$$

c) Calculate $\operatorname{Var}(N_{\boldsymbol{W}})$.
Solution. We have

$$\begin{aligned} \operatorname{Var}(N_{\boldsymbol{W}}) &= \operatorname{Var}\left(\sum_{i=1}^{M} \mathbf{1}(X_{j_i} = 1, X_{j_i+1} = 1)\right) \\ &= \sum_{i=1}^{M} \operatorname{Var}\left(\mathbf{1}(X_{j_i} = 1, X_{j_i+1} = 1)\right) \\ &\quad + \sum_{i \neq k} \operatorname{Cov}\left(\mathbf{1}(X_{j_i} = 1, X_{j_i+1} = 1), \mathbf{1}(X_{j_k} = 1, X_{j_k+1} = 1)\right) \\ &= \boxed{M(p^2 - p^4) + 2(M-1)(p^3 - p^4)}. \end{aligned}$$

NOTE: Due to time constraints, you can ignore the boundary effect.

2. Let T and C be independent Geometric random variables with succuess probability of r and s, respectively. That is,

$$P[T=j]=r(1-r)^{j-1}; \qquad j=1,2,\ldots,$$
$$P[C=j]=s(1-s)^{j-1}; \qquad j=1,2,\ldots.$$

Let $X=(\min(T,C), I(T\le C))$. Denote $X_1=\min(T,C)$ and $X_2=I(T\le C)$, where $I(\cdot)$ is the indicator function.

a) What is the joint distribution of X?
Solution. We have

$$\begin{aligned} P(X_1=k,X_2=1) &= P(\min(T,C)=k, T\le C)=P(T=k,C\ge k)=P(T=k)P(C\ge k)\\ &= r(1-r)^{k-1}\sum_{j=k}^{\infty} s(1-s)^{j-1} = r(1-r)^{k-1}s(1-s)^{k-1}\sum_{j=0}^{\infty}(1-s)^j\\ &= r(1-r)^{k-1}(1-s)^{k-1}, \end{aligned}$$

and

$$\begin{aligned} P(X_1=k,X_2=0) &= P(\min(T,C)=k, T> C)=P(C=k,T< k)=P(C=k)P(T> k)\\ &= s(1-s)^{k-1}\sum_{j=k+1}^{\infty} r(1-r)^{j-1} = s(1-s)^{k-1}(1-r)^k. \end{aligned}$$

Note that

$$P(X_1=k)=P(X_1=k,X_2=1)+P(X_1=k,X_2=0)+(r+s-rs)\left(1-(r+s-rs)\right)^{k-1}.$$

b) Calculate $\mathbb{E}X=(\mathbb{E}X_1,\mathbb{E}X_2)$ and the covariance matrix of $X=(X_1,X_2)$.
Solution. Since $X_1\sim \mathrm{Geom}(r+s-rs)$, we have

$$\mathbb{E}X_1=\frac{1}{r+s-rs}, \text{ and } \mathrm{Var}(X_1)=\frac{1-(r+s-rs)}{(r+s-rs)^2}.$$

Also,

$$\begin{aligned} \mathbb{E}X_2 = P(T\le C) &= \sum_{k=1}^{\infty} P(T\le k)P(C=k) = \sum_{k=1}^{\infty} s(1-s)^{k-1}\left(1-(1-r)^k\right)\\ &= \sum_{k=1}^{\infty} s(1-s)^{k-1} - \sum_{k=1}^{\infty} s(1-s)^{k-1}(1-r)^k = 1-\frac{s(1-r)}{r+s-rs} = \frac{r}{r+s-rs}, \end{aligned}$$

and so,

$$\mathrm{Var}(X_2)=\mathbb{E}X_2^2-(\mathbb{E}X_2)^2=\frac{r}{r+s-rs}-\frac{r^2}{(r+s-rs)^2}=\frac{rs(1-r)}{(r+s-rs)^2}.$$

As for the covariance, we see that

$$\begin{aligned} \mathbb{E}X_1X_2 &= \sum_{k=1}^{\infty} kP(X_1=k,X_2=1)=\sum_{k=1}^{\infty} kr(1-r)^{k-1}(1-s)^{k-1}\\ &= \frac{r}{r+s-rs}\sum_{k=1}^{\infty} k(r+s-rs)(1-r-s+rs)^{k-1} = \frac{r}{r+s-rs}\frac{1}{r+s-rs}=\frac{r^2}{r+s-rs}, \end{aligned}$$

where the last summation identity comes from the fact that the EV of $\mathrm{Geom}(p)$ is $1/p$. Hence,

$$\mathrm{Cov}(X_1,X_2)=\mathbb{E}X_1X_2-\mathbb{E}X_1-\mathbb{E}X_2=0,$$

and so, the covariance matrix is

$$\boxed{\frac{1}{(r+s-rs)^2}\begin{bmatrix} 1-r-s+rs & 0 \\ 0 & rs(1-r) \end{bmatrix}}.$$

c) Let $T_1, T_2, \ldots, T_n$ be a random sample from T, and $C_1, C_2, \ldots, C_n$ be a random sample from C. Define

$$S_1 = \sum_{i=1}^{n} \min(T_i, C_i)$$
$$S_2 = \sum_{i=1}^{n} I(T_i \leq C_i).$$

What is the maximum likelihood estimate $(\hat{r}, \hat{s})$ of (r, s), in terms of S_1 and S_2?

Solution. If s_2 is observed, without loss of generality, we may assume that it is the first s_2 such that $I(T_i \leq C_i) = 1$. Then, the likelihood function is

$$\begin{aligned}
\mathcal{L}(r, s; \boldsymbol{T}, \boldsymbol{C}) &= \prod_{i=1}^{s_2} P(T_i = t_i, C_i = 1) \prod_{i=s_2+1}^{n} P(T_i = t_i, C_i = 0) \\
&= \prod_{i=1}^{s_2} r\left[(1-r)(1-s)\right]^{t_i-1} \prod_{i=s_2+1}^{n} s(1-r)\left[(1-r)(1-s)\right]^{t_i-1} \\
&= r^{s_2}\left(s(1-r)\right)^{n-s_2}\left[(1-r)(1-s)\right]^{s_1-n},
\end{aligned}$$

and so, the log-likelihood function is

$$\log \mathcal{L}(r, s; \boldsymbol{T}, \boldsymbol{C}) = s_2 \log r + (n - s_2)\log s(1-r) + (s_1 - n)\log(1-r)\log(1-s).$$

The first partials are

$$\frac{\partial}{\partial r}\log\mathcal{L} = \frac{s_2}{r} - \frac{s_1 - s_2}{1-r}, \text{ and } \frac{\partial}{\partial s}\log\mathcal{L} = \frac{n-s_2}{s} - \frac{s_1-n}{1-s},$$

and setting them equal to zero gives

$$\hat{r} = \frac{S_2}{S_1}, \text{ and } \hat{s} = \frac{n - S_2}{S_1 - S_2}.$$

The second partials are

$$\frac{\partial^2}{\partial r^2}\log\mathcal{L} = -\frac{s_2}{r^2} - \frac{s_1-s_2}{(1-r)^2}, \frac{\partial^2}{\partial s^2}\log\mathcal{L} = -\frac{n-s_2}{s^2} - \frac{s_1-n}{(1-s)^2}, \text{ and } \frac{\partial^2}{\partial r \partial s}\log\mathcal{L} = 0.$$

Evaluating these at $\hat{r}$ and $\hat{s}$ gives us that the maximum of the likelihood is achieved, and so, the MLE of (r, s) is

$$\boxed{(\hat{r}, \hat{s}) = \left(\frac{S_2}{S_1}, \frac{n-S_2}{S_1-S_2}\right)}.$$