Stat S-110 Homework 5 Solutions, Summer 2015

The following problems are from Chapters 7-8 of the book.

1. (BH 7.2) Alice, Bob, and Carl arrange to meet for lunch on a certain day. They arrive independently at uniformly distributed times between 1 pm and 1:30 pm on that day.

(a) What is the probability that Carl arrives first?

For the rest of this problem, assume that Carl arrives first at 1:10 pm, and condition on this fact.

(b) What is the probability that Carl will have to wait more than 10 minutes for one of the others to show up? (So consider Carl's waiting time until at least one of the others has arrived.)

(c) What is the probability that Carl will have to wait more than 10 minutes for both of the others to show up? (So consider Carl's waiting time until both of the others has arrived.)

(d) What is the probability that the person who arrives second will have to wait more than 5 minutes for the third person to show up?

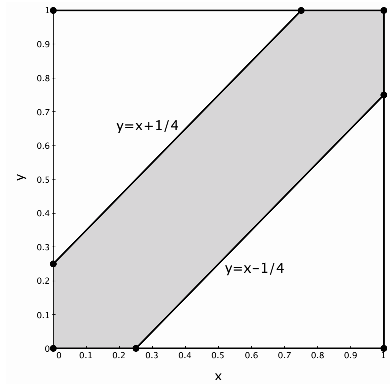
Solution:

(a) By symmetry, the probability that Carl arrives first is $1/3$.

(b) There is a 50% chance that Alice will arrive within the next 10 minutes and a 50% chance that Bob will arrive within the next 10 minutes. So by independence, the probability is $1/4$ that neither Alice nor Bob will arrive within the next 10 minutes.

(c) The probability is $1/4$ that both Alice and Bob will arrive within the next 10 minutes, so the probability is $3/4$ that Carl will have to wait more than 10 minutes in order for both Alice and Bob to have arrived.

(d) We need to find $P(|A-B| > 5)$, where A and B are i.id. Unif$(0, 20)$ r.v.s. Letting $X = A/20$ and $Y = B/20$, we need to find $P(|X - Y| > 0.25)$, where X and Y are i.i.d. Unif$(0, 1)$. This can be done geometrically, since the probability that (X, Y) is in a specific subregion of the square $\{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$ is proportional to the area of the subregion. In fact, the probability equals the area, since the area of the square is 1.



The region $|x-y| \leq 0.25$ is the shaded region in the square shown above. The complementary region, $|x-y| > 0.25$, consists of two disjoint triangles, each of which has area $\frac{1}{2}(\frac{3}{4})^2$. Therefore,

$$P(|X-Y| > 0.25) = \frac{9}{16}.$$

2. (BH 7.7) A stick of length L (a positive constant) is broken at a uniformly random point X. Given that $X = x$, another breakpoint Y is chosen uniformly on the interval $[0, x]$.

(a) Find the joint PDF of X and Y. Be sure to specify the support.

(b) We already know that the marginal distribution of X is $\text{Unif}(0, L)$. Check that marginalizing out Y from the joint PDF agrees that this is the marginal distribution of X.

(c) We already know that the conditional distribution of Y given $X = x$ is $\text{Unif}(0, x)$. Check that using the definition of conditional PDFs (in terms of joint and marginal PDFs) agrees that this is the conditional distribution of Y given $X = x$.

(d) Find the marginal PDF of Y.

(e) Find the conditional PDF of X given $Y = y$.

Solution:

(a) The joint PDF of X and Y is

$$f(x, y) = f_X(x) f_{Y|X}(y|x) = \frac{1}{Lx},$$

for $0 < x < L$ and $0 < y < x$ (and the joint PDF is 0 otherwise).

(b) Marginalizing out Y, while keeping in mind the constraint $y < x$, we have

$$f_X(x) = \int_0^x \frac{1}{Lx} dy = \frac{1}{L}$$

for $0 < x < L$, which agrees with the fact that $X \sim \text{Unif}(0, L)$.

(c) The conditional PDF of Y given $X = x$ is

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1/(Lx)}{1/L} = \frac{1}{x}$$

for $0 < y < x$, which agrees with the fact that $Y|X = x \sim \text{Unif}(0, x)$.

(d) The marginal PDF of Y is

$$f_Y(y) = \int_y^L f(x,y)dx = \frac{1}{L}\int_y^L \frac{1}{x}dx = \frac{\log L - \log y}{L},$$

for $0 < y < L$.

(e) The conditional PDF of X given $Y = y$ is

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{1}{x(\log L - \log y)},$$

for $y < x < L$.

3. (BH 7.8) (a) Five cards are randomly chosen from a standard deck, one at a time *with replacement*. Let X, Y, Z be the numbers of chosen queens, kings, and other cards. Find the joint PMF of X, Y, Z.

(b) Find the joint PMF of X and Y.

Hint: In summing the joint PMF of X, Y, Z over the possible values of Z, note that most terms are 0 because of the constraint that the number of chosen cards is five.

(c) Now assume instead that the sampling is without replacement (all 5-card hands are equally likely). Find the joint PMF of X, Y, Z.

Hint: Use the naive definition of probability.

Solution:

(a) By the story of the Multinomial, $(X, Y, Z) \sim \text{Mult}_3(5, \mathbf{p})$, with $\mathbf{p} = (1/13, 1/13, 11/13)$. So the joint PMF is

$$P(X = x, Y = y, Z = z) = \frac{5!}{x!y!z!} \cdot \left(\frac{1}{13}\right)^x \left(\frac{1}{13}\right)^y \left(\frac{11}{13}\right)^z = \frac{5!}{x!y!z!} \cdot \frac{11^z}{13^5},$$

for x, y, z nonnegative integers with $x + y + z = 5$.

(b) Since $X + Y + Z = 5$, the joint PMF of X and Y is

$$P(X = x, Y = y) = P(X = x, Y = y, Z = 5 - x - y) = \frac{5!}{x!y!(5-x-y)!} \cdot \frac{11^{5-x-y}}{13^5},$$

for x, y nonnegative integers with $x + y \leq 5$.

(c) By the naive definition of probability and the multiplication rule,

$$P(X = x, Y = y, Z = z) = \frac{\binom{4}{x}\binom{4}{y}\binom{44}{z}}{\binom{52}{5}},$$

for x, y, z nonnegative integers with $x + y + z = 5$.

4. (BH 7.46) Each of $n \geq 2$ people puts his or her name on a slip of paper (no two have the same name). The slips of paper are shuffled in a hat, and then each person draws one (uniformly at random at each stage, without replacement). Find the standard deviation of the number of people who draw their own names.

Solution: Label the people as $1, 2, \ldots, n$, let I_j be the indicator of person j getting his or her own name, and let $X = I_1 + \cdots + I_n$. By symmetry and linearity,

$$E(X) = nE(I_1) = n \cdot \frac{1}{n} = 1.$$

(This was also shown in the solution to Exercise 4.34.) To find the variance of X, we can expand in terms of covariances:

$$\begin{aligned}
\mathrm{Var}(X) &= n\mathrm{Var}(I_1) + 2\binom{n}{2}\mathrm{Cov}(I_1, I_2) \\
&= \frac{n}{n}\left(1 - \frac{1}{n}\right) + n(n-1)(E(I_1 I_2) - E(I_1)E(I_2)) \\
&= 1 - \frac{1}{n} + n(n-1)\left(\frac{1}{n(n-1)} - \frac{1}{n^2}\right) \\
&= 1 - \frac{1}{n} + 1 - \frac{n-1}{n} \\
&= 1.
\end{aligned}$$

Thus, the mean and standard deviation of X are both 1.

5. (BH 7.56) You are playing an exciting game of Battleship. Your opponent secretly positions ships on a 10 by 10 grid and you try to guess where the ships are. Each of your guesses is a *hit* if there is a ship there and a *miss* otherwise.

The game has just started and your opponent has 3 ships: a battleship (length 4), a submarine (length 3), and a destroyer (length 2). (Usually there are 5 ships to start, but to simplify the calculations we are considering 3 here.) You are playing a variation in which you unleash a *salvo*, making 5 simultaneous guesses. Assume that your 5 guesses are a simple random sample drawn from the 100 grid positions.

Find the mean and variance of the number of distinct ships you will hit in your salvo. (Give exact answers in terms of binomial coefficients or factorials, and also numerical values computed using a computer.)

Hint: First work in terms of the number of ships *missed*, expressing this as a sum of indicator r.v.s. Then use the fundamental bridge and naive definition of probability, which can be applied since all sets of 5 grid positions are equally likely.

Solution: Let N be the number of ships hit and M be the number of ships missed. Then $N + M = 3$, $E(N) = 3 - E(M)$, and $\text{Var}(N) = \text{Var}(3 - M) = \text{Var}(M)$. Write

$$M = I_2 + I_3 + I_4,$$

where I_j is the indicator of missing the ship of length j for $j \in \{2, 3, 4\}$. Then for $j, k \in \{2, 3, 4\}$ with $j \neq k$, by the fundamental bridge we have

$$E(I_j) = \frac{\binom{100-j}{5}}{\binom{100}{5}}, E(I_j I_k) = \frac{\binom{100-j-k}{5}}{\binom{100}{5}}.$$

Thus,

$$E(M) = E(I_2) + E(I_3) + E(I_4) = \frac{\binom{98}{5}}{\binom{100}{5}} + \frac{\binom{97}{5}}{\binom{100}{5}} + \frac{\binom{96}{5}}{\binom{100}{5}} \approx 2.57,$$

which gives $E(N) = 3 - E(M) \approx 0.43$. For the variance, we have

$$\begin{aligned} E(M^2) &= E(I_2^2) + E(I_3^2) + E(I_4^2) + 2E(I_2 I_3) + 2E(I_2 I_4) + 2E(I_3 I_4) \\ &= \frac{\binom{98}{5}}{\binom{100}{5}} + \frac{\binom{97}{5}}{\binom{100}{5}} + \frac{\binom{96}{5}}{\binom{100}{5}} + 2\left(\frac{\binom{95}{5}}{\binom{100}{5}} + \frac{\binom{94}{5}}{\binom{100}{5}} + \frac{\binom{93}{5}}{\binom{100}{5}}\right), \end{aligned}$$

so $\text{Var}(N) = \text{Var}(M) = E(M^2) - (E(M))^2$, with $E(M^2)$ and $E(M)$ as above. This evaluates to $\text{Var}(N) \approx 0.34$.

6. (BH 7.61) The number of people who visit the Leftorium store in a day is Pois(100). Suppose that 10% of customers are *sinister* (left-handed), and 90% are *dexterous* (right-handed). Half of the sinister customers make purchases, but only a third of the dexterous customers make purchases. The characteristics and behavior of people are independent, with probabilities as described in the previous two sentences. On a certain day, there are 42 people who arrive at the store but leave without making a purchase. Given this information, what is the conditional PMF of the number of customers on that day who make a purchase?

Solution: Let $N \sim \text{Pois}(100)$ be the number of people who visit the store in a day, X be the number who visit and make a purchase, and Y be the number who visit but don't make a purchase. So $X + Y = N$ and $X|N \sim \text{Bin}(N, p)$, where by LOTP

$$p = 0.1 \cdot \frac{1}{2} + 0.9 \cdot \frac{1}{3} = 0.35.$$

By the chicken-egg story, X and Y are independent, with $X \sim \text{Pois}(35), Y \sim \text{Pois}(65)$. So the conditional PMF of X given $Y = 42$ is

$$P(X = i|Y = 42) = P(X = i) = e^{-35} \cdot 35^i/i!,$$

for $i = 0, 1, 2, \ldots$.

7. (BH 7.63) There will be $X \sim \text{Pois}(\lambda)$ courses offered at a certain school next year.

(a) Find the expected number of choices of 4 courses (in terms of λ, fully simplified), assuming that simultaneous enrollment is allowed if there are time conflicts.

(b) Now suppose that simultaneous enrollment is not allowed. Suppose that most faculty only want to teach on Tuesdays and Thursdays, and most students only want to take courses that start at 10 am or later, and as a result there are only four possible time slots: 10 am, 11:30 am, 1 pm, 2:30 pm (each course meets Tuesday-Thursday for an hour and a half, starting at one of these times). Rather than trying to avoid major conflicts, the school schedules the courses completely randomly: after the list of courses for next year is determined, they randomly get assigned to time slots, independently and with probability 1/4 for each time slot.

Let X_{am} and X_{pm} be the number of morning and afternoon courses for next year, respectively (where "morning" means starting before noon). Find the joint PMF of X_{am} and X_{pm}, i.e., find $P(X_{\text{am}} = a, X_{\text{pm}} = b)$ for all a, b.

(c) Continuing as in (b), let X_1, X_2, X_3, X_4 be the number of 10 am, 11:30 am, 1 pm, 2:30 pm courses for next year, respectively. What is the joint distribution of

X_1, X_2, X_3, X_4? (The result is completely analogous to that of $X_{\text{am}}, X_{\text{pm}}$; you can derive it by thinking conditionally, but for this part you are also allowed to just use the fact that the result is analogous to that of (b).) Use this to find the expected number of choices of 4 non-conflicting courses (in terms of λ, fully simplified). What is the ratio of the expected value from (a) to this expected value?

Solution:

(a) Let k be the number of courses chosen (the problem is asking about the case $k = 4$, but it is not harder to solve this for general k). By LOTUS and the Taylor series for e^x,

$$E\binom{X}{k} = \sum_{n=k}^{\infty} \binom{n}{k} e^{-\lambda} \frac{\lambda^n}{n!} = \frac{e^{-\lambda}}{k!} \sum_{n=k}^{\infty} \frac{\lambda^n}{(n-k)!} = \frac{e^{-\lambda}\lambda^k}{k!} \sum_{n=k}^{\infty} \frac{\lambda^{n-k}}{(n-k)!} = \frac{\lambda^k}{k!}.$$

So for $k = 4$, there are $\frac{\lambda^4}{24}$ possibilities on average. As a check, note that when $k = 1$ the above reduces to $EX = \lambda$, and when $k = 2$ the result is true since $E(X(X-1)) = E(X^2) - EX = \text{Var}(X) + (EX)^2 - EX = \lambda^2$.

(b) This problem has exactly the structure of the chicken-egg problem, so X_{am} and X_{pm} are independent $\text{Pois}(\lambda/2)$ r.v.s. (Alternatively, condition on X, and use the fact that $P(X_{\text{am}} = a, X_{\text{pm}} = b | X = n) = 0$ if $a + b \neq n$; this is essentially how we solved the chicken-egg problem.) So the joint PMF is

$$P(X_{\text{am}} = a, X_{\text{pm}} = b) = \frac{e^{-\lambda/2}(\lambda/2)^a}{a!} \cdot \frac{e^{-\lambda/2}(\lambda/2)^b}{b!},$$

for all nonnegative integers a and b.

(c) Analogously to (b), X_1, X_2, X_3, X_4 are independent Poisson r.v.s. (A proof is given below, but was not required for this problem.) By symmetry and linearity, $E(X) = E(X_1) + E(X_2) + E(X_3) + E(X_4) = 4E(X_1)$, which gives $E(X_1) = \lambda/4$. So X_1, X_2, X_3, X_4 are independent $\text{Pois}(\lambda/4)$ r.v.s.

Thus, the average number of choices of course schedules if simultaneous enrollment is allowed is a factor of $256/24 = 32/3 \approx 10.67$ times larger than the average number of choices if simultaneous enrollment is not allowed.

8. (BH 7.69) Let X be the number of statistics majors in a certain college in the Class of 2030, viewed as an r.v. Each statistics major chooses between two tracks: a general track in statistical principles and methods, and a track in quantitative finance. Suppose that each statistics major chooses randomly which of these two tracks to follow, independently, with probability p of choosing the general track. Let

Y be the number of statistics majors who choose the general track, and Z be the number of statistics majors who choose the quantitative finance track.

(a) Suppose that $X \sim \text{Pois}(\lambda)$. (This isn't the exact distribution in reality since a Poisson is unbounded, but it may be a very good approximation.) Find the correlation between X and Y.

(b) Let n be the size of the Class of 2030, where n is a known constant. For this part and the next, instead of assuming that X is Poisson, assume that each of the n students chooses to be a statistics major with probability r, independently. Find the joint distribution of Y, Z, and the number of non-statistics majors, and their marginal distributions.

(c) Continuing as in (b), find the correlation between X and Y.

Solution:

(a) By the chicken-egg story, Y and Z are independent with $Y \sim \text{Pois}(\lambda p), Z \sim \text{Pois}(\lambda q)$, where $q = 1 - p$. So

$$\text{Cov}(X, Y) = \text{Cov}(Y + Z, Y) = \text{Cov}(Y, Y) + \text{Cov}(Z, Y) = \text{Var}(Y) = \lambda p.$$

Thus,

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\lambda p}{\sqrt{\lambda \lambda p}} = \sqrt{p}.$$

(b) We now have $X \sim \text{Bin}(n, r)$. Each of the n students becomes a Statistics concentrator in the General track with probability rp, a Statistics concentrator in the Quantitative Finance track with probability rq, and a non-Statistics concentrator with probability $1 - r$. By the story of the Multinomial,

$$(Y, Z, n - X) \sim \text{Mult}_3(n, (rp, rq, 1 - r)).$$

By the story of the Binomial, the marginal distributions are

$$Y \sim \text{Bin}(n, rp), Z \sim \text{Bin}(n, rq), n - X \sim \text{Bin}(n, 1 - r).$$

(c) By Theorem 7.4.6 (the result about covariances in a Multinomial),

$$\text{Cov}(X, Y) = \text{Cov}(Y + Z, Y) = \text{Var}(Y) + \text{Cov}(Y, Z) = nrp(1 - rp) - n(rp)(rq) = npr(1 - r).$$

So

$$\text{Corr}(X, Y) = \frac{npr(1 - r)}{\sqrt{nr(1 - r)nrp(1 - rp)}} = \sqrt{\frac{p(1 - r)}{1 - pr}}.$$

Note that if $n \to \infty$ and $r \to 0$ with nr fixed at a value λ, then $\text{Cov}(X, Y) \to \lambda p$, which is the covariance from (a); this makes sense since in this limit, the $\text{Bin}(n, r)$ distribution converges to the $\text{Pois}(\lambda)$ distribution.

9. (BH 8.18) Let X and Y be i.i.d. $\mathcal{N}(0, 1)$ r.v.s, and (R, θ) be the polar coordinates for the point (X, Y), so $X = R\cos\theta$ and $Y = R\sin\theta$ with $R \geq 0$ and $\theta \in [0, 2\pi)$. Find the joint PDF of R^2 and θ. Also find the marginal distributions of R^2 and θ, giving their names (and parameters) if they are distributions we have studied before.

Solution: We have $X = R\cos(\theta), Y = R\sin(\theta)$. Let $W = R^2, T = \theta$ and mirror the relationships between the capital letters via $w = r^2 = x^2 + y^2, x = \sqrt{w}\cos(t), y = \sqrt{w}\sin(t)$. By the change of variables formula,

$$f_{W,T}(w, t) = f_{X,Y}(x, y)\left|\frac{\partial(x, y)}{\partial(w, t)}\right| = \frac{e^{-r^2/2}}{2\pi}\left|\frac{\partial(x, y)}{\partial(w, t)}\right|.$$

The Jacobian matrix is

$$\frac{\partial(x, y)}{\partial(w, t)} = \begin{pmatrix} \frac{1}{2\sqrt{w}}\cos(t) & -\sqrt{w}\sin(t) \\ \frac{1}{2\sqrt{w}}\sin(t) & \sqrt{w}\cos(t) \end{pmatrix},$$

which has absolute determinant $\frac{1}{2}\cos^2(t) + \frac{1}{2}\sin^2(t) = \frac{1}{2}$. So the joint PDF of R^2 and θ is

$$f_{R^2,\theta}(w, t) = \frac{1}{4\pi}e^{-w/2} = \frac{1}{2\pi} \cdot \frac{1}{2}e^{-w/2},$$

for $w > 0$ and $0 \leq t < 2\pi$ (and 0 otherwise). Thus, R^2 and θ are independent, with

$$R^2 \sim \text{Expo}(1/2) \text{ and } \theta \sim \text{Unif}(0, 2\pi).$$

Note that this problem is Example 8.1.7 (Box-Muller) in reverse.

10. (BH 8.36) Alice walks into a post office with 2 clerks. Both clerks are in the midst of serving customers, but Alice is next in line. The clerk on the left takes an $\text{Expo}(\lambda_1)$ time to serve a customer, and the clerk on the right takes an $\text{Expo}(\lambda_2)$ time to serve a customer. Let T_1 be the time until the clerk on the left is done serving his or her current customer, and define T_2 likewise for the clerk on the right.

(a) If $\lambda_1 = \lambda_2$, is T_1/T_2 independent of $T_1 + T_2$?

Hint: $T_1/T_2 = (T_1/(T_1 + T_2))/(T_2/(T_1 + T_2))$.

(b) Find $P(T_1 < T_2)$ (do not assume $\lambda_1 = \lambda_2$ here or in the next part, but do check that your answers make sense in that special case).

(c) Find the expected total amount of time that Alice spends in the post office (assuming that she leaves immediately after she is done being served).

Solution:

(a) Let $W = T_1/(T_1+T_2)$. Then $T_1/T_2 = W/(1-W)$. By the bank–post office story, W is independent of T_1+T_2. So T_1/T_2 is also independent of T_1+T_2.

(b) We will show that $P(T_1 < T_2) = \lambda_1/(\lambda_1+\lambda_2)$. For $r > 0$, the CDF of T_1/T_2 at r is

$$
\begin{aligned}
P\left(\frac{T_1}{T_2} \le r\right) &= P(T_1 \le rT_2) \\
&= \int_0^\infty \left(\int_0^{rt_2} \lambda_1 e^{-\lambda_1 t_1} dt_1\right) \lambda_2 e^{-\lambda_2 t_2} dt_2 \\
&= \int_0^\infty (1 - e^{-\lambda_1 r t_2}) \lambda_2 e^{-\lambda_2 t_2} dt_2 \\
&= 1 - \int_0^\infty \lambda_2 e^{-(r\lambda_1+\lambda_2)t_2} dt_2 \\
&= 1 - \frac{\lambda_2}{r\lambda_1+\lambda_2} \\
&= \frac{r\lambda_1}{r\lambda_1+\lambda_2}.
\end{aligned}
$$

Letting $r = 1$, we have the claimed result.

In the special case $\lambda_1 = \lambda_2$, this gives $P(T_1 < T_2) = 1/2$, which we already knew to be true by symmetry.

(c) Alice's time in the post office is $T_{\text{wait}} + T_{\text{serve}}$, where $T_{\text{wait}} = \min(T_1, T_2)$ is how long she waits in line and T_{serve} is how long it takes for her to be served once it is her turn. By linearity, the expected total time is $E(T_{\text{wait}}) + E(T_{\text{serve}})$. Then

$$
\begin{aligned}
T_{\text{wait}} &\sim \text{Expo}(\lambda_1+\lambda_2), \\
E(T_{\text{wait}}) &= \frac{1}{\lambda_1+\lambda_2}.
\end{aligned}
$$

Let f_{serve} be the PDF of T_{serve}. By LOTP, conditioning on which clerk serves Alice,

$$
\begin{aligned}
f_{\text{serve}}(t) &= f_{\text{serve}}(t|T_1 < T_2)P(T_1 < T_2) + f_{\text{serve}}(t|T_1 > T_2)P(T_1 > T_2) \\
&= \lambda_1 e^{-\lambda_1 t} \cdot \frac{\lambda_1}{\lambda_1+\lambda_2} + \lambda_2 e^{-\lambda_2 t} \cdot \frac{\lambda_2}{\lambda_1+\lambda_2}.
\end{aligned}
$$

So

$$\begin{aligned}
E(T_{\text{serve}}) &= \int_0^\infty t f_{\text{serve}}(t)dt \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2}\int_0^\infty \lambda_1 t e^{-\lambda_1 t}dt + \frac{\lambda_2}{\lambda_1+\lambda_2}\int_0^\infty \lambda_2 t e^{-\lambda_2 t}dt \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2}\cdot\frac{1}{\lambda_1} + \frac{\lambda_2}{\lambda_1+\lambda_2}\cdot\frac{1}{\lambda_2} \\
&= \frac{2}{\lambda_1+\lambda_2}.
\end{aligned}$$

Thus, the expected total time is $3/(\lambda_1+\lambda_2)$.

11. (BH 8.40) An engineer is studying the reliability of a product by performing a sequence of n trials. Reliability is defined as the probability of success. In each trial, the product succeeds with probability p and fails with probability $1-p$. The trials are conditionally independent given p. Here p is unknown (else the study would be unnecessary!). The engineer takes a Bayesian approach, with $p \sim \text{Unif}(0,1)$ as prior.

Let r be a desired reliability level and c be the corresponding confidence level, in the sense that, given the data, the probability is c that the true reliability p is at least r. For example, if $r = 0.9, c = 0.95$, we can be 95% sure, given the data, that the product is at least 90% reliable. Suppose that it is observed that the product succeeds all n times. Find a simple equation for c as a function of r.

Solution: The prior distribution for p is $\text{Beta}(1,1)$, so the posterior distribution for p is $\text{Beta}(n+1,1)$. So

$$c = P(p \geq r|\text{data}) = (n+1)\int_r^1 p^n dp = 1 - r^{n+1}.$$

12. (BH 8.50) We are about to observe random variables $Y_1, Y_2, \ldots, Y_n$, i.i.d. from a continuous distribution. We will need to predict an independent future observation Y_{new}, which will also have the same distribution. The distribution is unknown, so we will construct our prediction using $Y_1, Y_2, \ldots, Y_n$ rather than the distribution of Y_{new}. In forming a prediction, we do not want to report only a single number; rather, we want to give a *predictive interval* with "high confidence" of containing Y_{new}. One approach to this is via order statistics.

(a) For fixed j and k with $1 \leq j < k \leq n$, find $P(Y_{\text{new}} \in [Y_{(j)}, Y_{(k)}])$.
Hint: By symmetry, all orderings of $Y_1, \ldots, Y_n, Y_{\text{new}}$ are equally likely.

(b) Let $n = 99$. Construct a predictive interval, as a function of $Y_1, \ldots, Y_n$, such that the probability of the interval containing Y_{new} is 0.95.

Solution:

(a) Using symmetry as in the hint, Y_{new} is equally likely to be in any of the $n+1$ possible "slots" relative to $Y_{(1)}, \ldots, Y_{(n)}$ (imagine inserting the new observation anywhere relative to the first n observations). There are $k - j$ slots that would put the new observation in between $Y_{(j)}$ and $Y_{(k)}$. So

$$P(Y_{\text{new}} \in [Y_{(j)}, Y_{(k)}]) = \frac{k-j}{n+1}.$$

(b) Take an interval $[Y_{(j)}, Y_{(k)}]$ with $k - j = 95$. For example, the interval $[Y_{(3)}, Y_{(98)}]$ is as desired.