

ACM116 - Fall 2006-2007 - Midterm Solutions

Handed out: 26 Oct 2006, Due: 31 Oct 2006

1/ A tribunal is enquiring on the paternity relationship between two people. For this it gives the analysis of blood phenotypes to two independent labs which give correct results with a probability α (lab A) and β (lab B). The probability that two people taken at random have the same phenotype is τ . Write I the event: the phenotypes are identical, A^+ (B^+) the event: Lab A (Lab B) finds that the phenotypes are identical. Making errors, is independent of the actual state of the patient.

- Compute $\mathbb{P}(A^+ \cap B^+ | I)$.
- Compute the probability that two people have the same phenotype knowing that the results of the two labs are positive.
- Compute the probabilities in parts a) and b) for the specific case $\alpha = \beta = 0.9, \tau = 10^{-3}$.

Solution:

- A^+ and B^+ are not independent, but conditioned on event I they are, so the events $\{A^+ | I\}$ and $\{B^+ | I\}$ are independent, and we can write

$$\mathbb{P}(A^+ \cap B^+ | I) = \mathbb{P}(A^+ | I) \mathbb{P}(B^+ | I) \quad (1)$$

$$= \alpha\beta. \quad (2)$$

- Use Bayes' Rule:

$$\mathbb{P}(I | A^+ \cap B^+) = \frac{\mathbb{P}(A^+ \cap B^+ | I) \mathbb{P}(I)}{\mathbb{P}(A^+ \cap B^+ | I) \mathbb{P}(I) + \mathbb{P}(A^+ \cap B^+ | I^c) \mathbb{P}(I^c)} \quad (3)$$

$$= \frac{\mathbb{P}(A^+ | I) \mathbb{P}(B^+ | I) \mathbb{P}(I)}{\mathbb{P}(A^+ | I) \mathbb{P}(B^+ | I) \mathbb{P}(I) + \mathbb{P}(A^+ | I^c) \mathbb{P}(B^+ | I^c) \mathbb{P}(I^c)} \quad (4)$$

$$= \frac{\alpha\beta\tau}{\alpha\beta\tau + (1-\alpha)(1-\beta)(1-\tau)}. \quad (5)$$

- Apply equation (5):

$$\mathbb{P}(I | A^+ \cap B^+) = 0.075. \quad (6)$$

With n labs, all returning results with accuracy α , we have

$$\mathbb{P}(I | A_1^+ \cap \dots \cap A_n^+) = \frac{1}{1 + \left(\frac{1-\alpha}{\alpha}\right)^n \left(\frac{1-\tau}{\tau}\right)}. \quad (7)$$

To achieve the reliable result $\mathbb{P}(I | A_1^+ \cap \dots \cap A_n^+) \approx 1$ we must have many accurate tests. In this case, it seems we have neither.

□

2/ Two weather stations are giving data on a climatic system which can be in two states S_1 and S_2 , shifting at random from one to the other. Long observations have shown that during 30% of the time the system is in the state S_1 and 70% of the time the system in the state S_2 . Station 1 gives erroneous

data in 2% of cases, and station 2 in 2% of cases. Making errors is independent of the actual state of the climatic system. Each station makes its errors independently of the other.

At a given time, station 1 is communicating that the system is in the state S_1 whereas station 2 is saying that the system is in the state S_2 . Which communication should be assumed to be correct?

Solution:

Let T_{11} be the event that station 1 reports state 1, and let T_{22} be the event that station 2 reports state 2. Apply Bayes' Rule:

$$\mathbb{P}(S_1|T_{11} \cap T_{22}) = \frac{\mathbb{P}(T_{11} \cap T_{22}|S_1)\mathbb{P}(S_1)}{\mathbb{P}(T_{11} \cap T_{22}|S_1)\mathbb{P}(S_1) + \mathbb{P}(T_{11} \cap T_{22}|S_2)\mathbb{P}(S_2)} \quad (8)$$

$$= \frac{\mathbb{P}(T_{11}|S_1)\mathbb{P}(T_{22}|S_1)\mathbb{P}(S_1)}{\mathbb{P}(T_{11}|S_1)\mathbb{P}(T_{22}|S_1)\mathbb{P}(S_1) + \mathbb{P}(T_{11}|S_2)\mathbb{P}(T_{22}|S_2)\mathbb{P}(S_2)} \quad (9)$$

$$= \frac{0.98 \times 0.02 \times 0.30}{0.98 \times 0.02 \times 0.30 + 0.02 \times 0.98 \times 0.70} \quad (10)$$

$$= 30\%. \quad (11)$$

With the given reports, the corresponding likelihood that the weather is in state S_2 is 70%, so we should believe station 2.

□

3/ It is commonly presumed that an unborn child has a 50% probability of being a female. But is it really the case? Let's look at birth statistics for the Netherlands for the years 1989, 1990 and 1991. According to the Central Bureau of Statistics, there were, in total 585609 children born during the span of those years, of which 286114 were girls. What is the estimate for the probability that a newborn child will be a girl and what is the corresponding 95% confidence interval. That is to say: writing p the probability that a newborn child will be a girl we are looking for a random variable \hat{p} given as a function of our statistics and a parameter $\alpha > 0$ such that

$$\mathbb{P}(p \in [\hat{p} - \alpha, \hat{p} + \alpha]) \approx 0.95, \quad (12)$$

where the randomness in \hat{p} corresponds to the randomness in our statistics. You may use the fact that

$$\int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \approx 0.95. \quad (13)$$

Solution:

Define the iid Bernoulli variables

$$X_i = \begin{cases} 1, & \text{child } i \text{ is a girl,} \\ 0, & \text{child } i \text{ is a boy.} \end{cases} \quad (14)$$

$$\mathbb{E}[X_i] = p, \quad (15)$$

$$\text{Var}(X_i) = p(1 - p). \quad (16)$$

Define

$$S_n = \sum_{i=1}^n X_i = \text{number of girls born in } n \text{ total births.} \quad (17)$$

Applying the Law of Large Numbers, we estimate for large n

$$S_n \approx n\mathbb{E}[X_i], \quad (18)$$

$$\mathbb{E}[X_i] \approx \frac{S_n}{n}, \quad (19)$$

$$\hat{p} = \frac{286114}{585609} \approx 0.4886. \quad (20)$$

The Central Limit Theorem gives that for sufficiently large n , $S_n \sim \mathcal{N}(np, np(1-p))$. Using the given definite integral,

$$0.95 \approx \mathbb{P} \left(-1.96 \leq \frac{S_n/n - p}{\sqrt{p(p-1)/n}} \leq 1.96 \right) \quad (21)$$

$$\approx \mathbb{P} \left(-1.96 \leq \frac{\hat{p} - p}{\sqrt{\hat{p}(\hat{p} - 1)/n}} \leq 1.96 \right), \quad (22)$$

where we have approximated the variance of X_i using \hat{p} instead of the unknown p . Thus, the half-width of the 95% confidence interval is approximately

$$\alpha = 1.96 \times \sqrt{\frac{\hat{p}(\hat{p} - 1)}{n}} \quad (23)$$

$$\approx 1.28 \times 10^{-3}. \quad (24)$$

and with 95% confidence,

$$p \in [0.4873, 0.4899]. \quad (25)$$

We have an alternative for estimating the variance of \hat{p} . Because we are testing the hypothesis that $p = 0.5$, we can in fact use $p = 0.5$ to compute the variance and then check if this choice satisfies our hypothesis. However, since \hat{p} is very close to 0.5, the half-width of the 95% confidence interval is still approximately 1.28×10^{-3} .

Strictly speaking, we should be using the Student's T distribution to represent the random variable \hat{p} . However, when n is large (and here it is enormous), the Student's T distribution converges to the Gaussian distribution.

In any case, we reach the conclusion that assuming p is constant across all populations and time periods, we can conclude with 95% confidence that there are fewer girls born than boys. Note that as n increases, our confidence interval narrows, as is to be expected.

□

4/ Suppose for instance that you are offered a sequence of bets, each bet being a losing proposition with probability $1 - p$ and paying out f times ($f > 1$) your stake with probability p .

- a) Compute the expected net payoff of each bet.
- b) Suppose that the expected net payoff of each bet is strictly positive (you have an edge). How to gamble if you must? The idea is to bet a fixed proportion of your present bankroll. When your bankroll decreases you bet less, as it increases you bet more. Assuming that your starting bankroll is V_0 , define the random variable V_n as the size of your bankroll after n bets when you bet a fixed fraction α ($0 < \alpha < 1$) of your current bankroll each time. Here it is supposed that winnings are reinvested and that your bankroll is infinitely divisible. Find the optimal value for α . Hint: Observe that $V_n = (1 - \alpha + \alpha R_1) \times \dots \times (1 - \alpha + \alpha R_n) V_0$ where R_k is equal to the payoff factor f if the k -th bet is won and is otherwise equal to 0.

Solution:

- a) Let s be the amount wagered on a single bet. The expected payoff is

$$\mathbb{E}[\text{Net Payoff}] = (f - 1)s \mathbb{P}(\text{Win}) - s \mathbb{P}(\text{Lose}) \quad (26)$$

$$= (f - 1)sp - s(1 - p) \quad (27)$$

$$= s(fp - 1). \quad (28)$$

Note that for the expected payoff to be positive, $fp - 1 > 0$. For example, if $p = 1/2$, we must have $f > 2$, as expected.

- b) **Method 1.** We assume that the bets, hence the R_k , are iid. Therefore, the factors $(1 - \alpha + \alpha R_k)$ are iid and we have

$$\mathbb{E}[V_n] = \mathbb{E} \left[V_0 \prod_{k=1}^n (1 - \alpha + \alpha R_k) \right] = V_0 \prod_{k=1}^n \mathbb{E}[1 - \alpha + \alpha R_k] \quad (29)$$

$$= V_0 (1 - \alpha + \alpha \mathbb{E}[R_1])^n = V_0 (1 + \alpha(fp - 1))^n \quad (30)$$

$$= V_0 (fp - 1)^n \left(\alpha + \frac{1}{fp - 1} \right)^n. \quad (31)$$

Since the expected net payoff for each bet is strictly positive, $fp - 1 > 0$, and on the interval $\alpha \in [0, 1]$, $\mathbb{E}[V_n]$ achieves its maximum at $\alpha = 1$. It seems that the optimal strategy is to bet the entire fortune on each bet.

This conclusion is insane, and **this method fails**. Why? Because if you play enough times, you will almost surely lose at least once, and when you do, you will lose your entire fortune. We have naïvely assumed that for large n , our fortune almost surely converges to its expectation, which is not in general true. What we must do is apply the Law of Large Numbers in a safe and orderly manner.

Method 2. Take the log of V_n/V_0 so that we can write our fortune after n trials as a sum of iid random variables:

$$\ln \frac{V_n}{V_0} = \sum_{k=1}^n \ln (1 - \alpha + \alpha R_k), \quad (32)$$

so by the Law of Large Numbers,

$$\frac{1}{n} \ln \frac{V_n}{V_0} \rightarrow \mathbb{E} [\ln (1 - \alpha + \alpha R_1)]. \quad (33)$$

$\ln(V_n/V_0)$ is an increasing function of V_n , so maximising $\mathbb{E} [\ln (1 - \alpha + \alpha R_1)]$ with respect to α maximises our fortune. Maximising

$$\mathbb{E} [\ln (1 - \alpha + \alpha R_1)] = p \ln(1 - \alpha + \alpha f) + (1 - p) \ln(1 - \alpha) \quad (34)$$

gives

$$\alpha = \frac{pf - 1}{f - 1}. \quad (35)$$

Since $pf > 1$, $f > 1$, and $p < 1$, we have $\alpha \in (0, 1)$. Only if $p = 1$ should we choose $\alpha = 1$.

□

- 5/ The waiting times at checkout desks of a supermarket are positive random variables X_0, \dots, X_n independent identically distributed with continuous density on $[0, \infty)$. Your waiting time is given by X_0 , and the waiting time of the person that arrived at the same time as you at the checkout desk i is X_i . Compute the law of $N = \inf\{i \geq 1; X_i > X_0\}$. Give its expectation. Observe that N is a measure of your frustration, i.e. a large N corresponds to a large number of people right next to you having checked before you.

Solution:

Method 1. The event $\{N = k\}$ for $1 \leq k \leq n$ is the event

$$\{X_1 \leq X_0, \dots, X_{k-1} \leq X_0, X_k > X_0\}. \quad (36)$$

The events $\{X_i \leq X_0\}$ are not independent, since they all depend on X_0 . The event $\{N = n + 1\}$ is the event

$$\{X_1 \leq X_0, \dots, X_n \leq X_0\}. \quad (37)$$

To compute the law of N , condition on X_0 . The events $\{X_i \leq X_0 | X_0 = x\}$ are independent, so for $1 \leq k \leq n$,

$$\mathbb{P}(N = k) = \mathbb{P}(X_1 \leq X_0, \dots, X_{k-1} \leq X_0, X_k > X_0) \quad (38)$$

$$= \int_0^\infty \mathbb{P}(X_1 \leq X_0, \dots, X_{k-1} \leq X_0, X_k > X_0 | X_0 = x) f_X(x) dx \quad (39)$$

$$= \int_0^\infty \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_{k-1} \leq x) \mathbb{P}(X_k > x) f_X(x) dx \quad (40)$$

$$= \int_0^\infty F_X(x)^{k-1} (1 - F_X(x)) f_X(x) dx, \quad (41)$$

$$= \int_0^\infty F_X(x)^{k-1} f_X(x) dx - \int_0^\infty F_X(x)^k f_X(x) dx, \quad (42)$$

where $F_X(x)$ and $f_X(x)$ are the cdf and pdf of the X_i , respectively. For the case $k = n + 1$,

$$\mathbb{P}(N = n + 1) = \mathbb{P}(X_1 \leq X_0, \dots, X_n \leq X_0) \quad (43)$$

$$= \int_0^\infty \mathbb{P}(X_1 \leq X_0, \dots, X_n \leq X_0 | X_0 = x) f_X(x) dx \quad (44)$$

$$= \int_0^\infty F_X(x)^n f_X(x) dx. \quad (45)$$

$\mathbb{P}(N = k) = 0$ for $\{k < 1, k > n + 1\}$; this fact together with equations (42) and (45) give the law of N for $1 \leq k \leq n + 1$.

The expectation of N is

$$\mathbb{E}[N] = \sum_{k=1}^{n+1} k \mathbb{P}(N = k) \quad (46)$$

$$= \int_0^\infty (1 - F_X(x)) f_X(x) dx + 2 \int_0^\infty F_X(x) (1 - F_X(x)) f_X(x) dx + \dots \quad (47)$$

$$+ n \int_0^\infty F_X(x)^{n-1} (1 - F_X(x)) f_X(x) dx + (n + 1) \int_0^\infty F_X(x)^n f_X(x) dx \quad (48)$$

$$= \int_0^\infty (1 + F_X(x) + F_X(x)^2 + \dots + F_X(x)^n) f_X(x) dx. \quad (49)$$

The integrals $\int_0^\infty F_X(x)^k f_X(x) dx$ do not depend of the law of X . Noting that $f_X(x) = F'_X(x)$, we make the substitution $u = F_X(x)$:

$$\int_0^\infty F_X(x)^k f_X(x) dx = \int_0^\infty F_X(x)^k F'_X(x) dx \quad (50)$$

$$= \int_{F_X(0)}^{F_X(\infty)} u^k du = \int_0^1 u^k du \quad (51)$$

$$= \frac{1}{1 + k} \quad (52)$$

Hence, the law of N and its expectation simplify significantly:

$$\mathbb{P}(N = k) = \begin{cases} \frac{1}{k(k+1)}, & k = 1, \dots, n, \\ \frac{1}{n+1}, & k = n + 1, \\ 0, & \text{otherwise,} \end{cases} \quad (53)$$

$$\mathbb{E}[N] = \sum_{k=1}^{n+1} \frac{1}{k}. \quad (54)$$

Since $\mathbb{E}[N] \rightarrow \infty$ from below as $n \rightarrow \infty$, under this model you might actually expect to be less frustrated at a supermarket with fewer checkouts.

Method 2. The drawback with method 1 is that the reason why the solution is independent of the specific choice for $F_X(x)$ is hidden in an integration. Instead, consider the following:

Which of the people is most likely to be served first? No one. They are all equally likely by symmetry. In fact, every possible monotonically increasing ordering of the X_i is equally likely, and all we are interested in is their ordering. Thus we could replace our problem by one that is easier to think about: We write the numbers $0, 1, 2, \dots, n$ on the backs of n cards. We shuffle the cards. Let Y_i be the value on the back of the i th card in their new random ordering. The original question is identical to examining the law of $N = \inf\{i \geq 1; Y_i > Y_0\}$.

$$\mathbb{P}(N = 1) = \mathbb{P}(\text{The second card is bigger than the first}) \quad (55)$$

$$= \frac{1}{2}. \quad (56)$$

This is a special case for all $1 \leq i \leq n$, where

$$\mathbb{P}(N = i) = \mathbb{P}(\text{Out of } i + 1 \text{ randomly ordered cards, the last one is the biggest} \\ \text{and the first one is the second biggest}) \quad (57)$$

$$= \frac{1}{i + 1} \cdot \frac{1}{i}. \quad (58)$$

Finally,

$$\mathbb{P}(N = n + 1) = \mathbb{P}(\text{The first card is the largest}) \quad (59)$$

$$= \frac{1}{n + 1}. \quad (60)$$

This gives the law of N . Now the expectation follows:

$$\mathbb{E}[N] = \left(\sum_{i=1}^n i \cdot \frac{1}{i(i+1)} \right) + (n+1) \cdot \frac{1}{n+1} \quad (61)$$

$$= \sum_{i=0}^n \frac{1}{i+1}. \quad (62)$$

This is identical to our solution from method 1, but here we see that the fact we only care about the ordering of the X_i is the reason that the law of N does not depend on the law of the X_i .

□