Mathematical Foundations of Biostatistics

Assignment 1

The assignment is due Monday September 30 by 11:59pm. All the questions have the same weight in the marking.

1. (a) Let $X_1 \sim \text{Binomial}(n_1, p)$, $X_2 \sim \text{Binomial}(n_2, p)$ and $X_1 \perp\!\!\!\perp X_2$. Find and name the conditional distribution of X_1 given $X_1 + X_2 = m$.

   (b) Let $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$ and $X_1 \perp\!\!\!\perp X_2$. Find and name the conditional distribution of X_1 given $X_1 + X_2 = m$.

   Recall here the binomial coefficient

   $$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

   where the factorial is given by $n! = n \times (n-1) \times \ldots \times 2 \times 1$.

2. Let X_1 and X_2 be independent and identically distributed random variables taking values in $\{-1, 1\}$ with

   $$P(X_i = 1) = 1 - P(X_i = -1) = p, \quad i = 1, 2.$$

   Let further $X_3 = X_1 X_2$.

   (a) Show that X_3 is a random variable.

   (b) Find the distribution of X_3 and $E[X_3]$. (Hint: note that there are two different ways of X_3 to receive a given value.)

   (c) Find $E[X_3 \mid X_i]$, $i = 1, 2$. Under which circumstances this does not depend on X_i?

   (d) Show that when $p = 0.5$ the random variables X_i and X_j are pairwise independent for all $i \neq j \in \{1, 2, 3\}$.

   (e) Show that when $p = 0.5$ the random variables X_1, X_2 and X_3 are not mutually independent.

Olli Saarela, Dalla Lana School of Public Health, 155 College Street, 6th floor, Toronto, Ontario M5T 3M7, Canada. E-mail: olli.saarela@utoronto.ca

(f) Show that when $p \neq 0.5$ the random variables X_3 and X_i, $i = 1, 2$ are not independent.

3. Bernstein (1924) discovered the inheritance pattern of ABO blood groups. Formerly, it was thought that the ABO bloodgroup phenotype was determined by two biallelic loci with alleles $\{A, a\}$ and $\{B, b\}$, respectively. Under this model, the individual has a bloodgroup phenotype 'AB' when both alleles A and B are present, bloodgroup 'A' when allele A is present and allele B is not present, bloodgroup 'B' when allele B is present and allele A is not present, and finally, bloodgroup 'O' when neither A or B is present. A and B are

   (a) List the possible genotypes under the two loci model (there are 9 in total).

   (b) In a population of 502 individuals, 42.2% had the observed bloodgroup phenotype 'A', 20.6% had the bloodgroup 'B', 7.8% had the bloodgroup 'AB', and the remaining 29.4% had the bloodgroup 'O'. Let us denote by $P(A = 1)$ the (marginal) probability that the allele A is present, and $P(B = 1)$ the (marginal) probability that the allele B is present. Using the observed data, and estimating the probabilities by the corresponding empirical proportions, calculate the probabilities $P(A = 1)$, $P(A = 0)$, $P(B = 1)$ and $P(B = 0)$. (Hint: use the law of total probability.)

   (c) Under the two loci model, and assuming the loci to be independent (not in linkage disequilibrium), calculate the probabilities of the four bloodgroup phenotypes $P(\text{'A'})$, $P(\text{'B'})$, $P(\text{'AB'})$ and $P(\text{'O'})$. (Hint: use the independence of the two loci and the marginal probabilities from (b).)

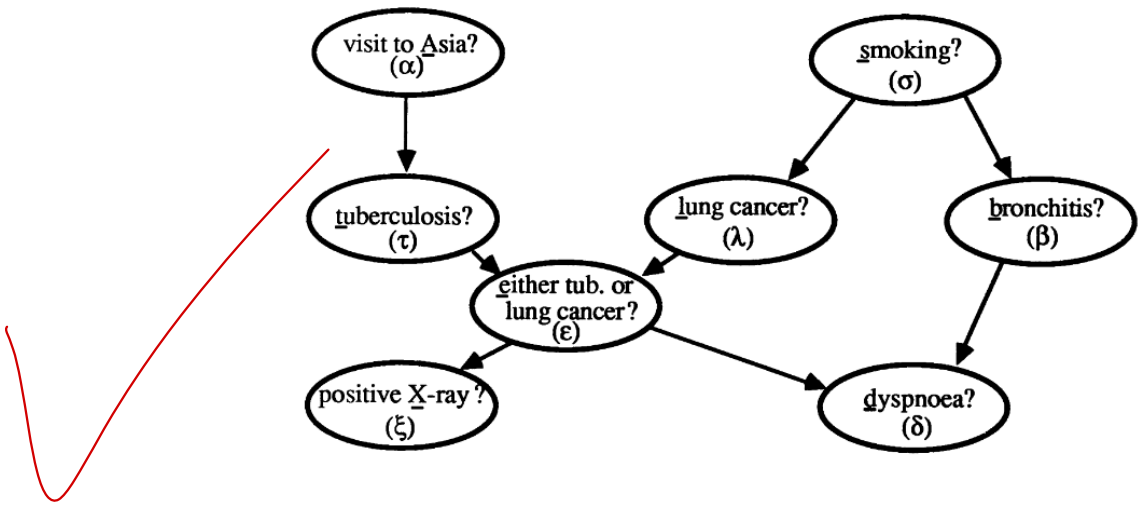
   (d) An alternative model is that the ABO bloodgroup is in fact determined by a single triallelic locus with the alleles $\{A, B, O\}$. List the possible genotypes under this model (there are 6 in total), and the bloodgroup phenotype related to each of these.

   (e) Suppose that from the observed data, the allele frequencies (that is, the probabilities that a copy of the particular allele is inherited from a given parent) under the single locus model were estimated as $P(A = 1) = 0.2945$, $P(B = 1) = 0.1547$ and $P(O = 1) = 0.5508$. Using these estimates, calculate the four

phenotype probabilities $P(\text{'AB'})$, $P(\text{'A'})$, $P(\text{'B'})$ and $P(\text{'O'})$ under the single locus model. (Hint: add the probabilities of different possible genotypes corresponding to a particular phenotype and assume that the maternal and paternal alleles are inherited independently).

(f) Which model fits better to the observed phenotype proportions? (Don't test, just compare the proportions.)

4. The below directed acyclig graph (DAG, from Lauritzen & Spiegelhalter 1988) represents a diagnostic problem concerning patients presenting at a chest clinic.



The corresponding dichotomous random variables are denoted as A, T, X, E, D, L, B, and S. DAGs are used to encode conditional indepence properties. In particular, the above DAG implies that $A \perp\!\!\!\perp L$ but $A \not\perp\!\!\!\perp L \mid X$. The interpretation of this is that given a positive X-ray, visiting Asia is informative of lung cancer status, even though these variables are marginally independent. DAGs also imply factorization of the joint distribution of the variables. In the present case, the implied factorization is

$$P(A,T,X,E,D,L,B,S)$$
$$= P(A)P(T \mid A)P(X \mid E)P(E \mid T,L)P(D \mid E,B)P(L \mid S)P(B \mid S)P(S).$$

Suppose that the conditional probabilities here were estimated as follows:

α:	$p(a)$	$= .01$	ε:	$p(e \mid l, t)$	$= 1$
				$p(e \mid l, \bar{t})$	$= 1$
τ:	$p(t \mid a)$	$= .05$		$p(e \mid \bar{l}, t)$	$= 1$
	$p(t \mid \bar{a})$	$= .01$		$p(e \mid \bar{l}, \bar{t})$	$= 0$
σ:	$p(s)$	$= .50$	ξ:	$p(x \mid e)$	$= .98$
				$p(x \mid \bar{e})$	$= .05$
λ:	$p(l \mid s)$	$= .10$			
	$p(l \mid \bar{s})$	$= .01$	δ:	$p(d \mid e, b)$	$= .90$
				$p(d \mid e, \bar{b})$	$= .70$
β:	$p(b \mid s)$	$= .60$		$p(d \mid \bar{e}, b)$	$= .80$
	$p(b \mid \bar{s})$	$= .30$		$p(d \mid \bar{e}, \bar{b})$	$= .10$

(a) Verify numerically that we have

$$\sum_{a,t,x,e,d,l,b,s} P(A = a, T = t, X = x, E = e, D = d, L = l, B = b, S = s) = 1.$$

(b) Verify numerically the result $A \perp\!\!\!\perp L$ by calculating the probabilities $P(L = 1)$, $P(L = 1 \mid A = 1)$, and $P(L = 1 \mid A = 0)$.

(c) Verify numerically the result $A \not\perp\!\!\!\perp L \mid X$ by calculating the probabilities $P(L = 1 \mid X = 1)$, $P(L = 1 \mid A = 1, X = 1)$, and $P(L = 1 \mid A = 0, X = 1)$.

1. (a) Let $X_1 \sim \text{Binomial}(n_1, p)$, $X_2 \sim \text{Binomial}(n_2, p)$ and $X_1 \perp\!\!\!\perp X_2$. Find and name the conditional distribution of X_1 given $X_1 + X_2 = m$.

   (b) Let $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$ and $X_1 \perp\!\!\!\perp X_2$. Find and name the conditional distribution of X_1 given $X_1 + X_2 = m$.

   Recall here the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

   where the factorial is given by $n! = n \times (n-1) \times \ldots \times 2 \times 1$.

a) $P(X_1 = k \mid X_1 + X_2 = m) = \dfrac{P(X_1 = k \text{ and } X_1 + X_2 = m)}{P(X_1 + X_2 = m)}$

$\overset{\text{ind.}}{=} \dfrac{P(X_1 = k)\, P(X_2 = m-k)}{P(X_1 + X_2 = m)}$

Note: $X_1 + X_2 \sim \text{Binomial}(n_1 + n_2, p)$

PMF of Binom: $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

$$= \frac{\left[\binom{n_1}{k} p^k (1-p)^{n_1-k}\right]\left[\binom{n_2}{m-k} p^{m-k} (1-p)^{n_2-(m-k)}\right]}{\binom{n_1+n_2}{m} p^m (1-p)^{n_1+n_2-m}}$$

$$= \frac{\binom{n_1}{k}\binom{n_2}{m-k} p^{k+m-k} \cancel{(1-p)^{n_1-k+n_2-(m-k)}}}{\binom{n_1+n_2}{m} \cancel{p^m} \cancel{(1-p)^{n_1+n_2-m}}}$$

$$= \frac{\binom{n_1}{k}\binom{n_2}{m-k}}{\binom{n_1+n_2}{m}}$$

population size, # of success, # of draws

$$\sim \text{Hypergeometric}(n_1+n_2,\ n_1,\ m)$$

b) $P(X_1 = k \mid X_1+X_2=m) = \dfrac{P(X_1=k \text{ and } X_1+X_2=m)}{P(X_1+X_2=m)}$

$$\overset{\text{ind.}}{=} \frac{P(X_1=k)\,P(X_2=m-k)}{P(X_1+X_2=m)}$$

Note: $X_1+X_2 \sim \text{Poisson}(\lambda_1+\lambda_2)$

PMF of Poisson: $P(X=x) = \dfrac{\lambda^x e^{-\lambda}}{x!}$

$$= \frac{\left[\dfrac{\lambda_1^{\,k} e^{-\lambda_1}}{k!}\right]\left[\dfrac{\lambda_2^{\,m-k} e^{-\lambda_2}}{(m-k)!}\right]}{\left[\dfrac{(\lambda_1+\lambda_2)^m e^{-(\lambda_1+\lambda_2)}}{m!}\right]}$$

$$= \frac{\lambda_1^{\,k}\,\lambda_2^{\,m-k}\,m!}{k!\,(m-k)!\,(\lambda_1+\lambda_2)^m}$$

$$= \frac{m!}{k!\,(m-k)!} \frac{\lambda_1^{\,k}\,\lambda_2^{\,m-k}}{(\lambda_1+\lambda_2)^{k+m-k}}$$

$$= \binom{m}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{m-k}$$

$$\sim \text{Binomial}\left(m, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$$

2. Let X_1 and X_2 be independent and identically distributed random variables taking values in $\{-1, 1\}$ with

$$P(X_i = 1) = 1 - P(X_i = -1) = p, \quad i = 1, 2.$$

Let further $X_3 = X_1 X_2$.

(a) Show that X_3 is a random variable.

(b) Find the distribution of X_3 and $E[X_3]$. (Hint: note that there are two different ways of X_3 to receive a given value.)

(c) Find $E[X_3 \mid X_i]$, $i = 1, 2$. Under which circumstances this does not depend on X_i?

(d) Show that when $p = 0.5$ the random variables X_i and X_j are pairwise independent for all $i \neq j \in \{1, 2, 3\}$.

(e) Show that when $p = 0.5$ the random variables X_1, X_2 and X_3 are not mutually independent.

(f) Show that when $p \neq 0.5$ the random variables X_3 and X_i, $i = 1, 2$ are not independent.

a) Since X_1, X_2 takes on $\{-1, 1\}$, consider $X_3 = X_1 X_2$

- If $X_1 = 1$, $X_2 = 1$ then $X_3 = 1$
- $X_1 = 1$, $X_2 = -1$ $X_3 = -1$
- $X_1 = -1$, $X_2 = 1$ $X_3 = -1$
- $X_1 = -1$, $X_2 = -1$ $X_4 = 1$

Hence, $X_3 \in \{-1, 1\}$.

Now consider $B = \{\{-1\}, \{1\}, \phi, \{-1, 1\}\}$, the σ-algebra associated with range of X_3. This can be verified trivially via the 3 axioms.

Next, we need to prove $X_3^{-1}(B) \in F$, the σ-algebra associated w/ Ω. This follows as X_1, X_2 are RVs themselves.

b) $P(X_3=1) = P(X_1=1)P(X_2=1) + P(X_1=-1)P(X_2=-1)$
$= p^2 + (1-p)^2$

$P(X_3=-1) = P(X_1=1)P(X_2=-1) + P(X_1=-1)P(X_2=1)$
$= p(1-p) + (1-p)p$
$= 2p(1-p)$

So distribution of X_3 is

X_3	1	-1
$P(X_3=x)$	$p^2+(1-p)^2$	$2p(1-p)$

$$E(X_3) = \sum_{x_i \in \{-1,1\}} x_i p(x_i) = (1)\left[p^2+(1-p)^2\right] + (-1)\left[2p(1-p)\right]$$
$$= p^2 + (1-p)^2 - 2p(1-p)$$
$$= 4p^2 - 4p + 1$$

c) Recall: $P(X_i=1)=p$ and $P(X_i=-1)=1-p$
Now $E(X_3|X_i) = E(X_1X_2|X_i)$ for $i=1,2$

1) $E(X_1X_2|X_1=1) = E(X_2) = p-(1-p) = 2p-1$
$E(X_1X_2|X_1=-1) = E(-X_2) = -p+(1-p) = -(2p-1)$
Hence, $E(X_3|X_1) = (2p-1)X_1$

2) $E(X_1X_2|X_2=1) = E(X_1) = p-(1-p) = 2p-1$

$E(X_1X_2 \mid X_2=-1) = E(-X_1) = -p+(1-p) = -(2p-1)$

Hence $E(X_3 \mid X_2) = (2p-1)X_2$

For X_3 not to depend on X_i, we need the term $(2p-1)X_2 = 0$

$\Rightarrow \quad 2p-1 = 0$

$\Rightarrow \quad p = 1/2$

(d) Show that when $p = 0.5$ the random variables X_i and X_j are pairwise independent for all $i \neq j \in \{1, 2, 3\}$.

(e) Show that when $p = 0.5$ the random variables X_1, X_2 and X_3 are not mutually independent.

---

d) Assume $P(X_i=1) = P(X_i=-1) = 1/2$ for $i \in \{1,2,3\}$.
We show for three cases: $X_1 \perp X_2$, $X_1 \perp X_3$, $X_2 \perp X_3$.

1) $X_1 \perp X_2$ by assumption.

2) Show $P(X_1X_3) = P(X_1)P(X_3)$.

First look for joint density

- If $X_1=1$, then $X_3 = X_1X_2 = X_2$. So
  $P(X_1=1, X_3=1) = P(X_1=1, X_2=1) = (0.5)^2 = 0.25$
  $P(X_1=1, X_3=-1) = P(X_1=1, X_2=-1) = (0.5)^2 = 0.25$
- If $X_1=-1$, then $X_3 = X_1X_2 = -X_2$. So
  $P(X_1=-1, X_3=1) = P(X_1=-1, X_2=-1) = (0.5)^2 = 0.25$
  $P(X_1=-1, X_3=-1) = P(X_1=-1, X_2=1) = (0.5)^2 = 0.25$

Now for marginal:

- $P(X_1=1) = P(X_1=-1) = 0.5$
- $P(X_3=1) = P(X_1X_2=1)$

$$\Leftrightarrow P(X_1 = X_2) = P(X_1=1)P(X_2=1) + P(X_1=-1)P(X_2=-1)$$
$$= (0.5)^2 + (0.5)^2$$
$$= 0.5$$

So $P(X_3=1) = 0.5$ ↑ similar reason

- $P(X_3=-1) = P(X_1X_2=-1) \Leftrightarrow P(X_1 \neq X_2) = \cdots = 0.5$

Lastly, checking the def^n of independence,

- $P(X_1=1, X_3=1) = 0.25 = P(X_1)P(X_3=1)$
- $P(X_1=1, X_3=-1) = 0.25 = P(X_1=1)P(X_3=-1)$
- $P(X_1=-1, X_3=1) = 0.25 = P(X_1=-1)P(X_3=1)$
- $P(X_1=-1, X_3=-1) = 0.25 = P(X_1=-1)P(X_3=-1)$

Hence $X_1 \perp X_3$

3) Using similar reasoning, it can be shown $X_2 \perp X_3$.

∎

e) Mutual independence: $P(X_1X_2X_3) = P(X_1)P(X_2)P(X_3)$.
We check for $X_1 = X_2 = X_3 = 1$.

Since $X_3 = X_1X_2$, then
$P(X_1=1, X_2=1, X_3=1) = P(X_1=1, X_2=1) = (0.5)^2 = 0.25$
But

$P(X_1=1)P(X_2=1)P(X_3=1) = (0.5)^3 = 0.125$

Hence, they are not mutually independent.

(f) Show that when $p \neq 0.5$ the random variables X_3 and X_i, $i = 1, 2$ are not independent.

WLOG we show $X_1 \not\perp X_3$.

Note $X_3 = X_1 X_2$, so now we check
$P(X_1=1, X_3=1) \neq P(X_1=1)P(X_3=1)$

LHS : $P(X_1=1, X_3=1) = P(X_1=1, X_2=1) = p^2$
RHS : $P(X_1=1)P(X_3=1) = p \times [p^2 + (1-p)^2]$

Clearly, LHS $\neq$ RHS when $p \neq 0.5$.

Similarly, we can say $X_2 \not\perp X_3$.

3. Bernstein (1924) discovered the inheritance pattern of ABO blood groups. Formerly, it was thought that the ABO bloodgroup phenotype was determined by two biallelic loci with alleles $\{A, a\}$ and $\{B, b\}$, respectively. Under this model, the individual has a bloodgroup phenotype 'AB' when both alleles A and B are present, bloodgroup 'A' when allele A is present and allele B is not present, bloodgroup 'B' when allele B is present and allele A is not present, and finally, bloodgroup 'O' when neither A or B is present.

   (a) List the possible genotypes under the two loci model (there are 9 in total).

   (b) In a population of 502 individuals, 42.2% had the observed bloodgroup phenotype 'A', 20.6% had the bloodgroup 'B', 7.8% had the bloodgroup 'AB', and the remaining 29.4% had the bloodgroup 'O'. Let us denote by $P(A = 1)$ the (marginal) probability that the allele A is present, and $P(B = 1)$ the (marginal) probability that the allele B is present. Using the observed data, and estimating the probabilities by the corresponding empirical proportions, calculate the probabilities $P(A = 1)$, $P(A = 0)$, $P(B = 1)$ and $P(B = 0)$. (Hint: use the law of total probability.)

a) AABB AABb AAbb
AaBB AaBb Aabb
aaBB aaBb aabb

b)

Blood Group	Prb
A	42.2
B	20.6
AB	7.8
O	29.4

$P(A=1) = P(A) + P(AB) = 0.422 + 0.078 = 0.5$

$P(A)$: Blood Type A; $P(AB)$: Type AB

$P(A=0) = 1 - 0.5 = 0.5$

$P(B=1) = P(B) + P(AB) = 0.206 + 0.078 = 0.284$

$P(B=0) = 1 - 0.284 = 0.716$

(c) Under the two loci model, and assuming the loci to be independent (not in linkage disequilibrium), calculate the probabilities of the four bloodgroup phenotypes $P(\text{'A'})$, $P(\text{'B'})$, $P(\text{'AB'})$ and $P(\text{'O'})$. (Hint: use the independence of the two loci and the marginal probabilities from (b).)

(d) An alternative model is that the ABO bloodgroup is in fact determined by a single triallelic locus with the alleles $\{A, B, O\}$. List the possible genotypes under this model (there are 6 in total), and the bloodgroup phenotype related to each of these.

c) $P(\text{'A'}) = P(A=1 \wedge B=0)$ ind.

$= P(A=1)\,P(B=0)$

$= (0.5)(0.716)$

$= 0.358$

$P(\text{'B'}) = P(A=0 \wedge B=1)$ ind.

$= P(A=0)\,P(B=1)$

$= (0.5)(0.284)$

$= 0.142$

$P(\text{'AB'}) = P(A=1 \wedge B=1)$ ind.

$= P(A=1)\,P(B=1)$

$= (0.5)(0.284)$

$= 0.142$

$$P(\text{'O'}) = P(A=O \wedge B=O)$$
$$= P(A=O)\,P(B=O)$$
$$= (0.5)(0.716)$$
$$= 0.358$$

d)

genotypes	bloodgroup phenotype
AA	A
BB	B
OO	O
AO	A
BO	B
AB	AB

(e) Suppose that from the observed data, the allele frequencies (that is, the probabilities that a copy of the particular allele is inherited from a given parent) under the single locus model were estimated as $P(A = 1) = 0.2945$, $P(B = 1) = 0.1547$ and $P(O = 1) = 0.5508$. Using these estimates, calculate the four phenotype probabilities $P(\text{'AB'})$, $P(\text{'A'})$, $P(\text{'B'})$ and $P(\text{'O'})$ under the single locus model. (Hint: add the probabilities of different possible genotypes corresponding to a particular phenotype and assume that the maternal and paternal alleles are inherited independently).

$$P(\text{'AB'}) = P(A=1)\,P(B=1) + P(B=1)\,P(A=1)$$
$$= 2\,P(A=1)\,P(B=1)$$
$$= 2\,(0.2945)(0.1547)$$
$$= 0.0911$$

$$P('A') = P(A=1)P(A=1) + 2P(A=1)P(O=1)$$
$$= (0.2945)^2 + 2(0.2945)(0.5508)$$
$$= 0.4111$$

$$P('B') = P(B=1)P(B=1) + 2P(B=1)P(O=1)$$
$$= (0.1547)^2 + 2(0.1547)(0.5508)$$
$$= 0.1943$$

$$P('O') = P(O=1)P(O=1)$$
$$= (0.5508)^2$$
$$= 0.3034$$

(f) Which model fits better to the observed phenotype proportions? (Don't test, just compare the proportions.)

Blood Grp	Observed	Two-loci Mod.	One-locus Mod.
A	42.2	35.8	41.4
B	20.6	14.2	19.4
AB	7.8	14.2	9.1
O	29.4	35.8	30.3

I think one-locus model is closer to the observed data.