BIOSTATISTICS (MATH-449)

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Mock exam

Date: 27th of May, 2025
Time: 10:15–12:00

Name: _____

SCIPER: _____

INSTRUCTIONS TO CANDIDATES

- This is a mock exam. The final one will contribute 80% to your final grade. To obtain the maximum number of points you should be clear about your reasoning and present your arguments explicitly. ~~You have **3 hours** to complete the exam.~~ You have **2 hours** to complete the mock exam. (We are not expecting to finish all questions within the 2 hours, but we wanted to give more opportunities to practice material.)
- All that can be used for this exam is a pen. No notes, books, summaries, formula collections or calculators are allowed. All questions should be answered.
- The finest enumerated item in each question will be marked on a scale of $0-2$ points, indicating an incorrect, partially correct and completely correct answer respectively (half-points are not given). **The mock exam has 4 questions with a total of 32 points**.
- **Write the answer to every question in *the other* booklet (Final exam - answers).** (This is a mock exam, you will not be provided with an answer booklet.) The questions are provided in this booklet. Scrap paper during the final exam will be provided for rough work, but only answers written in the other booklet will be marked.
- At the end of the exam, you will have to return everything: the booklet with the questions, the booklet with your answers, and the scrap paper.

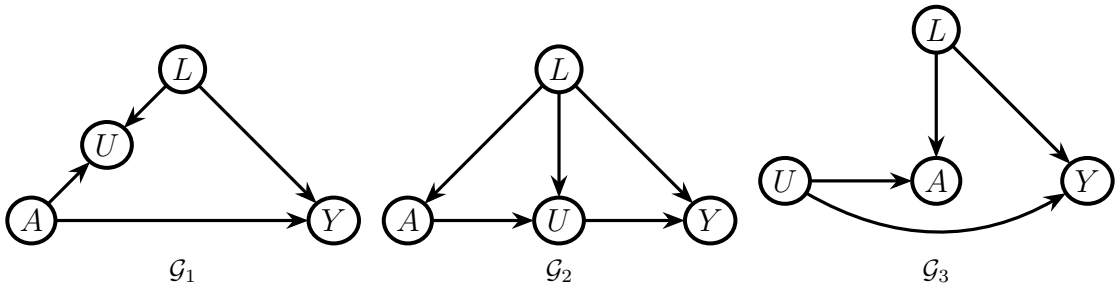
Mark question 1 (TOT: 8 points):

Mark question 2 (TOT: 6 points):

Mark question 3 (TOT: 6 points):

Mark question 4 (TOT: 12 points):

Question 1.
Let L, U, A and Y denote a measured baseline covariate, an unmeasured variable, treatment and outcome, respectively. Investigators 1, 2 and 3 propose causal models represented by the causal graphs $\mathcal{G}_1, \mathcal{G}_2$ and $\mathcal{G}_3$ respectively:



(a) Does any of the graphs $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ imply independencies in the observed law $p(l, a, y)$? Can the observed law $p(l, a, y)$ be used to falsify $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$? If yes, state the independencies that can be falsified.

(b) Consider the causal effect $E[Y^{a=1} - Y^{a=0}]$, assuming that consistency and positivity hold, and that
   (i) $\mathcal{G}_1$ is correct.
   (ii) $\mathcal{G}_2$ is correct.
   (iii) $\mathcal{G}_3$ is correct.
   For parts (b) (i)–(iii), express $E[Y^{a=1} - Y^{a=0}]$ as a function of the observed law $p(l, a, y)$ if this is possible, otherwise explain why it is not possible. Your expression should be "minimal", in the sense that none of the terms should contain unnecessary variables.

Solutions:

(a) $\mathcal{G}_1$ implies that $A \perp\!\!\!\perp L$. Therefore the observed law $p(l, a, y)$ can be used to falsify $\mathcal{G}_1$ whenever $p(l, a, y)$ violates $A \perp\!\!\!\perp L$.

Graphs $\mathcal{G}_2$ and $\mathcal{G}_3$ do not imply any independence in the observed law. Therefore, the observed law $p(l, a, y)$ cannot be used to falsify $\mathcal{G}_2$ or $\mathcal{G}_3$.

(b) (i) Under $\mathcal{G}_1$, $Y^a \perp\!\!\!\perp A$. Therefore,

$$E[Y^{a=1} - Y^{a=0}] = E[Y|A = 1] - E[Y|A = 0]$$

(ii) Under $\mathcal{G}_2$, $Y^a \perp\!\!\!\perp A|L$. Therefore,

$$E[Y^{a=1} - Y^{a=0}] = \sum_l \left(E[Y|A = 1, L = l] - E[Y|A = 0, L = l]\right) P(L = l)$$

(iii) Under $\mathcal{G}_3$, we have neither $Y^a \perp\!\!\!\perp A$, nor $Y^a \perp\!\!\!\perp A|L$. However, there is no causal effect of A on Y. Thus, $Y^{a=1} = Y^{a=0} = Y$ and $E[Y^{a=1} - Y^{a=0}] = 0$

Question 2. (a) Let $\alpha(t)$ be a hazard function and $S(t)$ the corresponding survival function for a (continuously distributed) survival time T. Show the identity

$$S(t) = \exp\left(-\int_0^t \alpha(s)\, ds\right).$$

(b) Let $Z_j(t)$ denote the number of individuals at risk and $N_j(t)$ the counting processes of events in group j. State the formula for the Kaplan-Meier estimators of the survival functions for the groups $j = 1, 2$.

(c) It is sometimes considered insufficient to just present Kaplan-Meier (survival) curves without a formal test of the difference between the groups. The most commonly used test for this purpose is the log-rank test, which is based on a test statistic defined on a time interval $[0, \tau]$. This statistic can be expressed as

$$Q(\tau) = N_2(\tau) - \int_0^\tau \frac{Z_2(t)}{Z_\bullet(t)} dN_\bullet(t),$$

where $Z_\bullet(t) = Z_1(t) + Z_2(t)$ and $N_\bullet(t) = N_1(t) + N_2(t)$.

Show that, under the null hypothesis of equal survival distributions in the two groups, it holds that $E(Q(\tau)) = 0$.

Hint: You can use that, under the null hypothesis, the counting process $N_j(t)$ had cumulative intensity process $\Lambda_j(t) = \int_0^t Z_j(s)\alpha(s)\, ds$ where $\alpha(s)$ is the common hazard function under the null hypothesis.

Solutions:

(a) By definition the hazard $\alpha(t) = \frac{f(t)}{S(t)} = -\frac{dS(t)}{S(t)}$. Thus, the cumulative hazard equals $\Lambda(t) = \int_0^t \lambda(s)\,ds = -\int_0^t \frac{dS(s)}{S(s)} = -\log(S(t))$ giving $S(t) = \exp\left(-\int_0^t \alpha(s)\,ds\right)$.

(b) The Kaplan-Meier estimators are given as

$$\hat{S}(t) = \prod_{s \le t} \left[1 - \frac{dN_j(s)}{Z_j(s)}\right].$$

(c) We can rewrite hazards for the groups, since under the null $\alpha_1(t) = \alpha_2(t) = \alpha(t)$,

$$\begin{aligned} Z(\tau) &= N_2(\tau) - \int_0^\tau \frac{Z_2(t)}{Z_\bullet(t)} dN_\bullet(t) \\ &= \int_0^\tau \frac{Z_1(t)}{Z_\bullet(t)} dN_2(t) - \int_0^\tau \frac{Z_2(t)}{Z_\bullet(t)} dN_1(t) \\ = \left(\int_0^\tau \frac{Z_1(t)}{Z_\bullet(t)} d\Lambda_2(t) - \int_0^\tau \frac{Z_2(t)}{Z_\bullet(t)} d\Lambda_1(t)\right) &- \left(\int_0^\tau \frac{Z_1(t)}{Z_\bullet(t)} dM_2(t) - \int_0^\tau \frac{Z_2(t)}{Z_\bullet(t)} dM_1(t)\right) \\ = \left(\int_0^\tau \frac{Z_1(t)Z_2(t)}{Z_\bullet(t)} \alpha(t)\,dt - \int_0^\tau \frac{Z_2(t)Z_1(t)}{Z_\bullet(t)} \alpha(t)\,dt\right) &+ M^*(\tau) = 0 + M^*(\tau) \end{aligned}$$

where $M^*(t) = -\left(\int_0^t \frac{Z_1(s)}{Z_\bullet(s)} dM_2(s) - \int_0^t \frac{Z_2(s)}{Z_\bullet(s)} dM_1(s)\right)$ is a martingale with expectation zero. Thus $E(Z(\tau)) = 0$ under the null hypothesis.

Question 3. We will consider the so-called accelerated failure time (AFT) model. In this model, the distribution of a survival time T conditional on a covariate x is given through the relation

$$\log(T) = \mu - \beta' x + \sigma W$$

where β' is a regression parameter, W is a random variable typically with mean zero, and μ and σ are so-called location and dispersion parameters generating a family of distributions $\mu + \sigma W$.

(a) Show that the survival function for T given x can be expressed as

$$S(t \mid x) = S_0(\exp(\beta' x)t)$$

where $S_0(t) = P(\exp(\mu + \sigma W) > t)$.

Give an explanation for the name "AFT-model" based on this representation (1-3 sentences).

(b) Show that the hazard function of T given x can be written as $\alpha(t \mid x) = \exp(\beta' x)\alpha_0(\exp(\beta' x)t)$, where $\alpha_0(t) = \alpha(t \mid 0) = -\frac{d}{dt}\log(S_0(t))$ is the hazard function when $x = 0$.

(c) Assume now that $S_0(t) = \exp(-bt^k)$ for parameters $b > 0$ and $k > 0$, which means that $\exp(\mu + \sigma W)$ has a Weibull distribution (you do not need to know more about this distribution). Demonstrate that, in this case, the accelerated failure time model is also a proportional hazards model on the form $\alpha(t \mid x) = h_0(t)\exp(\gamma' x)$ with a constant hazard ratio.

Determine the relation between the coefficient β' in the accelerated failure time model and the coefficient γ' in the proportional hazards model.

Solutions:

(a) We have

$$\begin{aligned} S(t \mid x) &= P(T > t) = P(\log(T) > \log(t)) = P(\mu - \beta' x + \sigma W > \log(t)) \\ &= P(\mu + \sigma W > \log(t) + \beta' x) = P(\exp(\mu + \sigma W) > t \exp(\beta' x)) \\ &= S_0(\exp(\beta' x)t). \end{aligned}$$

This survival function depends on time t multiplied by an acceleration factor $\exp(\beta' x)$, so we can consider time as moving $\exp(\beta' x)$ faster with covariate x.

(b) To get to the hazard function, we use the representation from Problem 1, question (a): $S(t) = \exp\left(-\int_0^t \alpha(s)\, ds\right)$ which corresponds to

$$\alpha(t) = \frac{d}{dt}(-\log(S(t))) = \frac{-dS(t)}{S(t)}.$$

Hence, the hazard in an AFT model is given as

$$\alpha(t \mid x) = \frac{\exp(\beta' x)(-S_0'(\exp(\beta' x)t))}{S_0(\exp(\beta' x)t)} = \exp(\beta' x)\alpha_0(\exp(\beta' x)t).$$

(c) We can use the first representation $S(t \mid x) = S_0(\exp(\beta' x)t)$, which with $S_0(t) = \exp(-bt^k)$ gives

$$S(t \mid x) = \exp(-b(\exp(\beta' x)t)^k) = \exp(-b(\exp(k\beta' x)t^k)) = \exp(-b(\exp(\gamma' x)t^k)),$$

which is also the survival function of a Weibull distribution. The hazard corresponding to this survival function is given as

$$\alpha(t \mid x) = bk \exp(\gamma' x)t^{k-1} = \exp(\gamma' x)bkt^{k-1} = \exp(\gamma' x)h_0(t),$$

where $h_0(t) = bkt^{k-1}$ is the baseline corresponding to a so-called Weibull survival function $S_0(t) = \exp(-bt^k)$. We recognize this model as a proportional hazards model with proportionality factor $\exp(\gamma' x) = \exp(k\beta' x)$. Thus the regression parameter in the proportional hazard model equals $\gamma = k\beta$ compared to the model $\log(T) = \mu - \beta' x + \sigma W$.

Question 4. Consider a chain-binomial SIR model with discrete time dynamics, for a fixed population of size n. Infections are assumed to happen in (discrete) generations. In the following, we assume that once an individual becomes infective, they remain infective for one generation, and then they are immunized, hence, they are removed. Moreover, each susceptible form generation t remains susceptible in generation $t+1$, if they avoided infection from all infectives of generation t. Infectious events at a given generation t occur independently from each other. Denote the number of susceptibles and infectives in generation t with S_t and I_t, respectively.

(a) In the *Reed-Frost model*, it is assumed that each infective has the same per contact infection probability p in every generation. Moreover, each potential infection is not affected by any other contacts, neither from previous generations nor in the current generation. That is, if a susceptible individual escaped infections in the previous generations, or was not infected by other contacts in the same generation, they did not develop immunity, nor did they become more susceptible. Under these assumptions, write down the transition probability of the (Reed-Frost) model, that is, derive

$$P(S_{t+1} = s_{t+1}, I_{t+1} = i_{t+1} | S_0 = s_0, \dots, S_t = s_t, I_0 = i = 0, \dots I_t = i_t).$$

Does $\{S_t, I_t\}_{t=0,1,\dots}$ satisfy the Markov property? Justify!

(b) Suppose you observe a particular chain of the number of infectives $\{i_0, i_1, \dots i_T = 0\}$. Write down the probability of observing this chain, in other words, what is

$$P(I_0 = i_0, \dots, I_T = 0)?$$

(c) Denote the final number of infected among the initially susceptible with Z, and the probability of observing $Z = j$ given $S_0 = k, I_0 = i$ with

$$P(Z = j | S_0 = k, I_0 = i) =: m_{ijk}.$$

Argue that the recursive expression

$$m_{ijk} = \binom{k}{j} m_{ijj} q^{(i+j)(k-j)}, \quad j < k, \quad \text{where } m_{ikk} = 1 - \sum_{j=0}^{k-1} m_{ijk},\ q = 1 - p,$$

is correct. A derivation of the formula is sufficient but not necessary for obtaining full points.

(d) Suppose that the number of infectives in the new generation, does not depend on the exact number of the infectives in the previous generation, just on the presence of at least one infective. That is, if there is any infective present in a generation, the probability of infecting a given individual is p if there is any infective present, otherwise it is 0. Under this new assumption, corresponding to the *Greenwood model*, write down the transition probability

$$P(S_{t+1} = s_{t+1}, I_{t+1} = i_{t+1} | S_0 = s_0, \dots, S_t = s_t, I_0 = i = 0, \dots I_t = i_t).$$

(e) Give a formula for the reproduction number R_0 both in the Reed-Frost and the Greenwood model. Is it true that in these models, if $R_0 > 1$ the epidemic will always *take off*, that is, a large outbreak will occur? Justify!

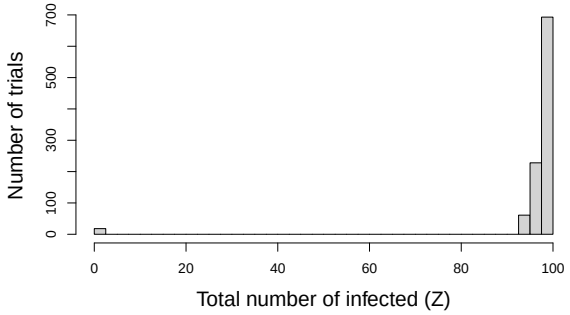
(f) We have simulated $N = 1000$ realizations from both the Reed-Frost and the Greenwood models, with fixed parameters and initial conditions, $I_0 = 1$, $n = 100$, under

different infection probabilities p. The distribution of the final size of the epidemic, can be estimated with the histograms, depicted below.

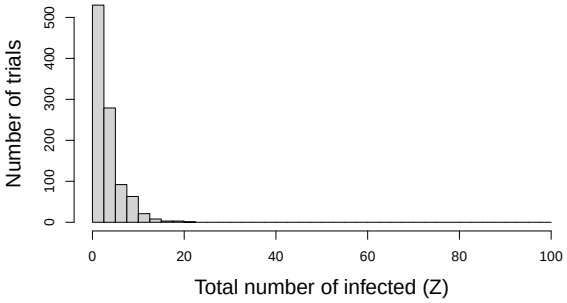
Match the different model and infection probability pairs to the histograms.

(i) $p = 1 \cdot 10^{-2}$, *Reed-Frost*
(ii) $p = 2 \cdot 10^{-2}$, *Reed-Frost*
(iii) $p = 4 \cdot 10^{-2}$, *Reed-Frost*
(iv) $p = 1 \cdot 10^{-2}$, *Greenwood*
(v) $p = 4 \cdot 10^{-2}$, *Greenwood*

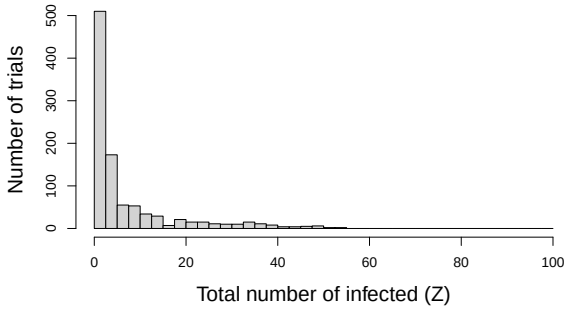
Give a brief justification (3-4 sentences in total) for your choice.



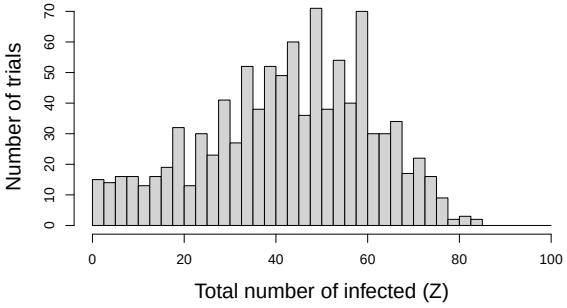
(A)



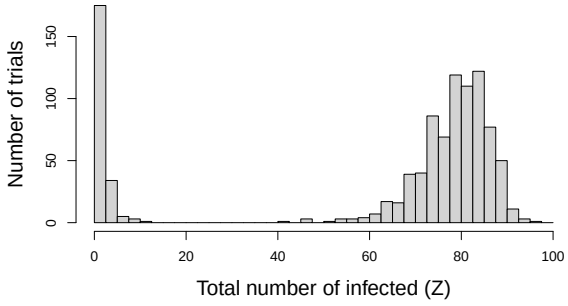
(B)



(C)



(D)



(E)

Solutions:

(a) Based on the description of the model, the number of new infectives in the new generation only depends on the number of infectives and susceptibles in the previous generation. In addition, the number of new susceptibles is deterministically follows from the number of new infectives and the number of previous susceptibles, that is

$$P(S_{t+1} = s_{t+1}, I_{t+1} = i_{t+1} | S_0 = s_0, \dots, S_t = s_t, I_0 = i = 0, \dots I_t = i_t)$$
$$= \begin{cases} P(I_{t+1} = i_{t+1} | S_t = s_t, I_t = i_t) & \text{if } s_{t+1} = s_t - i_{t+1} \\ 0 & \text{otherwise,} \end{cases}$$

therefore the Markov property is satisfied.

Since each susceptible individual can avoid infection by a given infectious individual with probability $1-p$ independently from each other, the probability, that a susceptible avoids infection, conditional on that there are i_t many infective in generation t is $(1-p)^{i_t}$, so the probability that they get infected is $1-(1-p)^{i_t}$. Since we condition on that there are s_t many susceptibles in generation t, and infections (e.g. successes) happen independently from each other, the new infections have a Binomial$(s_t, (1-(1-p)^{i_t}))$ distribution, thus the transition probability is

$$P(I_{t+1} = i_{t+1} | S_t = s_t, I_t = i_t) = \binom{s_t}{i_{t+1}} \left(1-(1-p)^{i_t}\right)^{i_{t+1}} \left((1-p)^{i_t}\right)^{s_t - i_{t+1}}.$$

(b) Using that there exists a deterministic relationship between the susceptibles and the infectives, that is, $s_0 = n - i_0$, $s_1 = s_0 - i_1$, $s_2 = s_1 - i_2$, ...

$$P(I_0 = i_0, \dots, I_T = 0) = P(I_0 = i_0, \dots, I_T = 0, S_0 = s_0, \dots, S_{T-1} = s_{T-1}),$$

for the recursively determined values of s_t. Then, by the definition of conditional probability and the fact that $\{I_t, S_t\}$ is Markov, we can rewrite

$$\begin{aligned} &P(I_0 = i_0, \dots, I_T = 0, S_0 = s_0, \dots, S_T = s_T) \\ &\quad = P(I_T = 0, S_T = s_T | I_0 = i_0, \dots, I_{T-1} = i_{t-1}, S_0 = s_0, \dots, S_{T-1} = s_{T-1}) \\ &\quad \cdot P(I_{T-1} = i_{T-1}, S_{T-1} = s_{T-1} | I_0 = i_0, \dots, I_{T-2} = i_{t-2}, S_0 = s_0, \dots, S_{T-2} = s_{T-2}) \dots \\ &\quad \cdot P(I_0 = i_0, S_0 = s_0) \\ &\quad = P(I_T = 0 | I_{T-1} = i_{T_1}, S_{T-1} = s_{T-1}) \cdot P(I_{T-1} | I_{T-2} = i_{T-2}, S_{T-2} = s_{T-2}) \dots \\ &\quad \cdot P(I_1 | I_0 = i_0, S_0 = s_0) \cdot 1 \\ &\quad = \prod_{t_0}^{T-1} \binom{s_t}{i_{t+1}} \left(1-(1-p)^{i_t}\right)^{i_{t+1}} \left((1-p)^{i_t}\right)^{s_t - i_{t+1}}. \end{aligned}$$

(c) m_{ijk} stands for the probability of observing an infectious chain that had in total j many new infectives among the initially k many susceptibles, if we had initial conditions $S_0 = k$, $I_0 = i$. Denote, the probability, that precisely the set $\mathcal{K} \subseteq \{1, \dots k\}$ with $|\mathcal{K}| = j$ was infected during the course of the epidemic with $m_{i\mathcal{K}k}$. Due to symmetry

$$m_{ijk} = \binom{k}{j} m_{i\mathcal{K}k}.$$

Divide the total susceptible population into two compartments, $\mathcal{K}$ and $\mathcal{K}^c = \{1, \ldots, k\} \setminus \mathcal{K}$. We can consider the final size probabilities in these two compartments separately. For $\mathcal{K}$, all susceptibles within this group must be infected, that is, we start from $j = |\mathcal{K}|$ susceptibles and i infectives, and we are interested in the probability that all j susceptibles will become infected during the epidemic, that corresponds to m_{ijj}.

For $\mathcal{K}^c$, none of the initial susceptibles can become infected. That is, in the first generation, they should avoid infection by the i many initial susceptibles, with probability $(q^i)^{k-j}$. Then in the subsequent generations, they should escape infection from all newly generated infectives, until all individuals in $\mathcal{K}$ are infected. This can happen with probability

$$(q^{i_1})^{k-j}(q^{i_2})^{k-j} \ldots (q^{i_{T_\mathcal{K}}})^{k-j},$$

where $T_\mathcal{K}$ is the index of the generation when the last individuals in $\mathcal{K}$ became infected. Since we fixed that, only the individuals in $\mathcal{K}$ became infected, and that $|\mathcal{K}| = j$, it follows that

$$(q^{i_1})^{k-j}(q^{i_2})^{k-j} \ldots (q^{i_{T_\mathcal{K}}})^{k-j} = (q^j)^{k-j}.$$

Putting everything together and using the independence of infectious contacts

$$m_{ijk} = \binom{k}{j} m_{i\mathcal{K}k} = \binom{k}{j} m_{ijj}(q^i)^{k-j}(q^j)^{k-j} = \binom{k}{j} m_{ijj}(q^{i+j})^{k-j},$$

which gives the desired result.

(d) We can use an analogous argument as in the solution of sub-question (a), except that now, given that there are $i_t > 0$ many infective in generation t, the probability that a given susceptible becomes infected is p and the escape probability is $1 - p$. Thus the number of new infections is Binomial($\mathrm{s_t}$, p) distributed, that is

$$P(I_{t+1} = i_{t+1} | S_t = s_t, I_t = i_t) = \binom{s_t}{i_{t+1}} p^{i_{t+1}} q^{s_t - i_{t+1}} \text{ if } s_t \geq i_{t+1}, i_t > 0,$$

and it is 0 otherwise.

(e) The reproduction number denotes the expected number of infections induced by a single infective individual in a large, completely susceptible population. In the population of size n, the population cannot be "larger" than n. Thus, for both the Reed-Frost and the Greenwood model, the expected number of infections induced by a single infectious individual is $n \cdot p \approx s_0 \cdot p$.

Since we are in the context of *stochastic* epidemic models, there is a non-zero probability, that the infection will die out before "blowing up". For example, the probability that the infection will die out after the first generation is $(q^{i_t})^{s_t}$, that is non-zero, even when q is small and s_t is large, i.e. $R_0 >> 1$.

(f) The pairs are (iii, A), (iv, B), (i, C), (v, D), (ii, E).

The lowest infection probability ($1 \cdot 10^{-2} \implies R_0 = 1$) corresponds to figures B and C, as in those cases, the epidemic dies out with only a small number of susceptibles being infected. In the case of the Greenwood model, it does not matter if $i_1 >> 1$, the probability of extinction in the next generation is the same regardless of the current infective size, while in case of the Reed-Frost model, once the size of the infectives start to increase, the infection is less likely to go extinct. Thus, the distribution

under the Reed-Frost model has a longer tail/larger variance, so Reed-Frost is C and Greenwood is B.

By similar reasoning, in the case of the Reed-Frost model, if $R_0 > 1$, either the infection dies out quickly, or there is a possibility of a major outbreak, since as the number of infectives increases, the probability of new infections increases exponentially. Therefore, sub-figures A and E, correspond to the Reed-Frost model, and since the expected final size of the epidemic is greater for sub-figure A, it corresponds to $p = 4 \cdot 10^{-2}$ and B to $p = 2 \cdot 10^{-2}$.

By all other options being excluded, sub-figure D corresponds to the Greenwood model with $p = 4 \cdot 10^{-2}$. Alternatively, since the probability of new infections (successes) are independent of the current size of the infectious population, if the infection takes off, the distribution of the Greenwood model is "Binomial-like" (it is not Binomial, since the probability of new infections still depend on the current size of the susceptible population).