Variáveis aleatórias: Projetos breves

O solicitado para cada projeto é:

- Preparar uma apresentação de dez minutos respondendo perguntas do enunciado (não se espera que respondam todas).
- A apresentação deve descrever o problema (que é diferente de ler o enunciado) e embasar teoricamente os métodos utilizados para responder as perguntas (numéricos ou analíticos). Se recomenda utilizar cálculos analíticos, nos casos que seja possível, para verificar que a programação esteja correta.
- Para a primeira parte do projeto (que será apresentada como **Prova 4**) o solicitado é realizar uma série de experimentos aleatórios do sistema de interesse de maneira a gerar um conjunto de dados simulados (ou sintéticos). A seguir, aplicar técnicas de análise descritiva para responder as perguntas e descrever o comportamento do sistema.
- Sintam-se livres para pensar alguma variante que vocês achem interessante. Algumas estão sugeridas em alguns enunciados. Podem adicionar uma pergunta original sua, algum estudo que gere curiosidade ou surpresa.
- Cada alun@ (não cada grupo) deverá encaminhar os slides da apresentação e os códigos desenvolvidos pelo e-disciplinas num arquivo zip (mesmo que seja um pdf, por favor zipar). Serão avaliados os seguintes aspectos: (a) Clareza e precisão, (b) correção dos cálculos, raciocínios, códigos e resultados, (c) criatividade do conteúdo original, (d) qualidade dos gráficos.
- A pessoa do grupo que realizará a apresentação será sorteada no momento da prova. Todos devem ser capazes de apresentar, responder perguntas, e rodar os códigos utilizados se for pedido.
- **Se espera que cada grupo tenha no mínimo uma reunião preparatória com um monitor ou docente, antes das apresentações de (que serão a Prova 4).**

1. Confiabilidade

Um sistema requer n máquinas para funcionar. Elas quebram, a cada unidade de tempo (um minuto), com probabilidade p_0. Para garantir o funcionamento confiável do sistema, s máquinas adicionais são mantidas para reposição. Quando uma máquina quebra, ela é substituída por uma em bom estado e enviada para reparo. O tempo de reparo de cada máquina é t_R (fixo).

Deseja-se estudar a variável aleatória "tempo de colapso" do sistema, isto é, o tempo T que o sistema funciona até dever ser parado porque não há n máquinas disponíveis. Quanto vale $E(T)$? Qual é a distribuição de T? Como depende a variável T do número de máquinas reserva s? Mais em concreto, quantas máquinas adicionais são necessárias para que Prob$\{T < 10000\}$ seja menor que 1%?

Faixas sugeridas de valores: $n = 10 - 100$, $p_0 = 0.001 - 0.01$, $t_R = 10 - 100$.

- Quais das perguntas acima conseguem responder analiticamente? Nos casos que possa, qual é o resultado?
- Os resultados do simulador são consistentes com as estimações analíticas no caso $\beta = 0$?
- No simulador, é fácil inserir a amostragem de outras variáveis aleatórias. Estudar por exemplo a variável Z que podemos llamar "tempo de alerta", definido como o tempo em que o número de máquinas em reparação atinge 80% do valor disponível s (ou outra porcentagem a escolher). Como é a distribuição de probabilidade de Z?
- São Z e T variáveis independentes? É possível predizer em alguma medida o tempo de colapso uma vez que foi atingido o tempo de alerta? Se uma parada ordenada do sistema requer o tempo t_R, pode sugerir uma estratégia de manejo que "garanta" chegar a parada ordenada com 99% de probabilidade?
- É lógico que as máquinas envelheçam. A probabilidade de cada máquina de falhar pode crescer linearmente com o tempo, seguindo $p(k) = p_0 + \beta\, t_f(k)$ onde k identifica a máquina e $t_f(k)$ o tempo que ela leva em funcionamento desde o último reparo. No caso de $\beta \neq 0$ é impossível realizar cálculos analíticos sobre a variável T. Para isto, desenvolver um pequeno código de simulação do processo. Esse caso é mais complexo que o original, porque é necessário armazenar informação específica para cada máquina (o tempo que leva trabalhando). Mesmo assim, trata-se apenas de uma adequada combinação de variáveis Bernoulli que o computador sabe simular.

2. Preço vs. Falência

Um pescador vende peixe fresco grelhado numa barraca na praia. Ele recebe o pedido, pesca, prepara o peixe, grelha, serve e vende. Ele nunca armazena mais de um peixe. Cada hora ele checa se apareceu um novo pedido, o que acontece com probabilidade p (valor típico: 0.4, consideramos que no máximo chega um pedido por hora), e decide o que fazer na próxima hora. Se ele tem um cliente e um peixe, a próxima hora é dedicada a cozinhar o peixe, servir e receber o preço S. Se ele tem cliente mas não tem peixe, ele vai pescar, com probabilidade q de conseguir um peixe (valor típico: 0.7). Se ele tem peixe e não tem cliente, ele descansa. Se ele não tem cliente nem peixe, ele vai pescar. O custo de manutenção da barraca é de 10 reais por hora. Qual é o preço que ele deve vender o peixe para o negócio não ir a falência, dependendo dos valores de p e q?

Nada é livre de risco. Vamos assumir que o pescador decide determinar o preço de maneira que seu risco de ir a falência nas primeiras 900 horas de funcionamento seja menor que 1%. É considerada falência quando as perdas acumuladas são maiores que o capital inicial do pescador.

- Considerar $p = 0.4 - 0.7$, $q = 0.6 - 0.9$ e capital inicial de 500 a 2000 reais. Um pescador com maior capital inicial poderia vender mais barato e levar a falência a outro com menor capital?
- *Sugestão:* Considerar que o pescador tem 4 estados possíveis: (1) Sem cliente nem peixe, (2) sem cliente com peixe, (3) com cliente e sem peixe, (4) com cliente e com peixe. Ele irá de um estado a outro (ou não) hora por hora, com probabilidades fáceis de calcular. Se ele a tempo t está no estado 4, por exemplo, ele com probabilidade 1 estará no estado 1 a tempo $t+1$, e seu capital terá crescido em $S-10$ reais.
- Uma variável aleatória X que nos interessa é o mínimo, entre $t=0$ e $t=900$, do capital $C(t)$ ao tempo t, sendo que $C(0)$ = capital inicial. Cada realização com $X<0$ corresponde a uma falência.
- Notar que, para cada $t>0$, o capital $C(t)$ é uma v.a. diferente (mas não independente) de $C(t-1)$, $C(t-2)$, etc.
- Escolha o preço S de acordo com o critério do pescador. Qual é o capital final esperado, após 900 horas, condicionado a não ter falido? Escrita matematicamente, a pergunta é qual é a distribuição da v.a. $C(t=900)$ condicionado a $X \geq 0$.
- Analisando o processo do pescador, existem pedidos não atendidos? Adicionar uma v.a. Y que seja o número de pedidos não atendidos em 900 horas.

- Seria interessante para o pescador saber quantas tempo horas seguidas ele terá que trabalhar, entre dois descansos.
- Poderia acontecer que o esforço colocado na pesca, e assim a probabilidade q de pescar, não fossem fixas. Imaginemos por exemplo que se, ao tempo t, o capital $C(t)$ é maior que o capital inicial, o pescador relaxa um pouco seus métodos e com isto a probabilidade de pesca cai para $0.8q$. Nesse caso, qual seria o valor apropriado de S?

3. Caminhada aleatória em 2D

Um móvel sai da origem (0,0), sendo que apenas pode se movimentar por uma rede x-y de pontos de distanciamento 1. Assim, a cada passo de tempo, da posição (i, j) pode ir a $(i+1, j)$, $(i-1, j)$, $(i, j+1)$ e $(i, j-1)$. O móvel realiza esse movimento repetidamente, de maneira que ao longo do experimento vai passando pelas posições $(x(t), y(t)) \in \mathbb{Z}^2$, com $t > 0$.

O movimento se inicia com igual probabilidade nas 4 direções, e a partir do segundo movimento, as probabilidades são: $1/4+\alpha$ de repetir a direção do passo anterior, $1/4-\alpha$ de voltar atrás, e $1/4$ de ir em cada uma das outras duas direções. Poderia ser visto como uma espécie de inercia do processo. Quando α é zero temos a caminhada aleatória "padrão".

O móvel está dentro da caixa $-L \leq x, y \leq L$, e registramos o tempo T que ele demora em chegar à borda da caixa, e qual é o ponto da borda (X_C, Y_C), onde faz contato e é removido.

- Nos interessa analisar as distribuições das variáveis T, X_C e Y_C. Tem partes da borda onde os móveis demoram mais em chegar? Como é a distribuição das "idades" T dos móveis que chegam aos diversos pontos da borda da caixa?
- Essas distribuições dependem de α? Em quais variáveis é observada maior dependência?
- Como dependem as variáveis de interesse do tamanho L?
- Quanto tempo deveria durar o experimento para ter 95% de probabilidade de o móvel ter chegado à borda?

4. Uma epidemia de Poisson

Um modelo básico de epidemia numa população de N indivíduos considera eles divididos em quatro grupos: S susceptíveis (sãos mas sem imunidade), I infectados não hospitalizados, H infectados hospitalizados e R recuperados (com imunidade). Esses números evoluem, considerando intervalos δt de tempo, segundo as seguintes regras:

- Se $I(t)+H(t)$ e $S(t)$ são o número de infectados e susceptíveis, respectivamente, ao tempo t, produzem-se novos contágios com uma taxa média $r_I(t) = \beta(t)\,(I(t) + H(t))S(t)/N$ (em contágios/dia). Desses novos contágios, 5% em média é internado no hospital. Se sabe que normalmente a população tem $\beta(t) = \beta_0 = 0.2592$ (contágios/dia), e que isto depende da frequência dos contatos inter-pessoais. Para certos dias excepcionais (cuja frequência média é de 2 por mês) acontecem grandes aglomerações levando o valor de β a $3\beta_0$.
- A taxa média de recuperação dos infectados não hospitalizados (em pacientes por unidade de tempo) é modelada como $r_{RI}(t) = \gamma_I\, I(t)$, com $\gamma_I = 0.07143$ dia^{-1} (duração média da doença $\simeq$ duas semanas).
- A taxa média de recuperação dos infectados hospitalizados é modelada como $r_{RH}(t) = \gamma_H\, H(t)$, com $\gamma = 0.03571$ dia^{-1} (duração média da doença $\simeq$ quatro semanas).

Deseja-se construir um simulador desse processo.

Considere $N = 10^4$ e o número inicial de infectados $I(0)$ aleatório, entre 10 e 20. Inicialmente não há infectados hospitalizados. O tempo a ser estudado é até não haver mais pessoas hospitalizadas por 3 dias seguidos. As variáveis de interesse são, dentre outras:

- O máximo número de hospitalizados simultâneos, isto é, $X = \max_t H(t)$.
- O tempo T_* desde a primeira hospitalização até o pico maximo de hospitalizados.
- A fração da população que teve a doença uma vez acabada a epidemia. A fração que esteve hospitalizada.
- O tempo desde a primeira hospitalização até a hospitalização número 10, e até a hospitalização número 40 (essas variáveis denotaremos por T_{10} e T_{40}). Notar que essas variáveis são fáceis de observar numa epidemia real.

- Sugestão de modelo de Poisson:

```
p = dt*beta(t)*(I(t-1)+H(t-1))*S(t-1)/N;
novosI   = randp(0.95*p);
novosH   = randp(0.05*p);
qI = dt*gammaI*I(t-1);
recuperadosI = randp(qI)
qH = dt*gammaH*H(t-1);
recuperadosH = randp(qH);
novosRecuperados = recuperadosI + recuperadosH;
```

- A distribuição de probabilidade de T_* é de bastante interesse. Qual seria o valor de t tal que Prob$(T_* < t) = 0.05$?
- Como variam os resultados do modelo quando é variada a frequência dos eventos de aglomeração para 1/mês ou 3/mês? Como variam se a taxa de internação não é 5% (por exemplo, se fosse 1%)?
- Os resultados do modelo são razoavelmente independentes do passo de tempo δt utilizado? Deveriam?
- É possível utilizar T_{10} e T_{40} para estimar em que momento acontece o pico de hospitalizações?

5. Uma enquête eleitoral

Sabe-se que a população de um pais está dividida em dois grupos políticos, que chamaremos A e B. A fração do grupo A é f, a de B é $1 - f$. Três perguntas serão feitas numa enquête, com respostas X, Y e Z, números reais. O pessoal de A acostuma responder às três perguntas com $N(0, \sigma)$, e o pessoal de B com $N(1, \sigma)$.

- Como são as distribuições de X, Y e Z para diversos valores de σ e de f? E as distribuições conjuntas?
- Para uma certa resposta (x_i, y_i, z_i), é possível estimar a qual dos grupos o respondente pertence?
- Suponha que se extrai uma amostra de n pessoas, cujo grupo político se desconhece. O grupo de cada pessoa é estimado segundo se $x + y + z$ é maior ou menor que $3/2$. Dessa maneira é calculada a "fração estimada", sendo
$$f_{\text{est}} = \frac{1}{n} \sum_i \chi\left(x_i + y_i + z_i < \frac{3}{2}\right)$$
onde χ é 1 se o argumento é verdadeiro, e 0 se não.
- f_{est} é uma v.a., cujo espaço amostral é a de todas as possíveis amostras de n pessoas. Como é sua distribuição de probabilidade? Como depende de n, de σ e de f?

- Vários dos itens colocados acima podem ser calculados analiticamente. Realizem o cálculo e um estudo de simulação, e vejam se há concordância entre os resultados.
- É possível calcular um valor de n tal que o erro $|f_{\text{est}} - f|$ seja menor que 1%? Verificar com simulações.