

Parte XIV

Técnicas estatísticas para análise de dados e de resultados de modelos de simulação

- A saída de um modelo de simulação geralmente constitui-se de VA's, muitas das quais podem ter variância grande.
 - Portanto, não é seguro fazer inferências (tomar decisões) com base na saída de uma única execução do modelo.
- Seja Y_1, Y_2, \dots a saída de um processo estocástico gerado por uma execução da simulação.
 - Ex.: Y_i pode representar o número de itens produzidos na i -ésima hora.
 - Geralmente os Y_i 's são VA que não são independentes nem identicamente distribuídas.
 - Portanto, os métodos clássicos para determinação dos intervalos de confiança não podem ser **diretamente** aplicados.
- Agora assuma que vamos executar a simulação n vezes, utilizando o mesmo estado inicial.
 - Se coletarmos m observações de cada simulação, teremos no total nm observações $y_{j,i}$ (a i -ésima observação da execução j).
 - Para cada execução, vamos utilizar conjuntos independentes de números aleatórios. Portanto, mesmo que as observações de uma mesma execução j não sejam IID, as i -ésimas observações de todas as execuções são IID (para qualquer i).
 - Portanto, o objetivo da análise de resultados é utilizar as observações $y_{j,i}$ para realizar inferências sobre as distribuições das VA's Y_1, \dots, Y_m .
 - Ex.: $\bar{y}_i = \sum_{j=1}^n y_{j,i}/n$ é um estimador não viesado de $E[Y_i]$.

48 Comportamento transiente e estacionário de um processo estocástico

- Considere o processo estocástico Y_1, Y_2, \dots
 - Seja $F_i(y|C) = \Pr[Y_i \leq y | C]$, $i = 1, 2, \dots$, onde y é um número real e C representa as condições utilizadas no início da simulação (no instante 0).

- Ex.: a condição inicial em um sistema de filas pode ser o tamanho de cada fila.
- $F_i(y|C)$ é chamado de “distribuição transiente” da saída de um processo no instante (discreto) i , para as condições iniciais C .
- Geralmente, $F_i(y|C)$ é diferente para cada i (Figura 9.1, Law).

- Para valores fixos de y e C , se $F_i(y|C) \rightarrow F(y)$ quando $i \rightarrow \infty$ (para qualquer escolha de y e C), dizemos que $F(y)$ é a “distribuição estacionária” do processo Y_1, Y_2, \dots

- A distribuição estacionária não depende das condições iniciais C , mas a taxa de convergência depende (Figura 9.2).

49 Tipos de simulação com relação à análise de dados

- A análise de dados mais apropriada depende de algumas características da simulação.
 - Vamos então classificar a simulação de acordo com estas características.
- Uma simulação é “finita” se existe um evento “natural” E que determina a duração de uma execução.
 - O evento E normalmente ocorre em um instante a partir do qual nenhuma informação relevante pode ser produzida, ou quando o sistema é reinicializado.
 - Ex.: para analisar o tempo de espera de um cliente em um banco, sabemos que os clientes serão atendidos se chegarem entre 10h e 16h. Depois de 16h nenhum novo cliente pode entrar, e os que já estavam no banco serão atendidos. A condição inicial é de zero clientes.
 - Ex.: determinar qual tropa vai vencer em uma simulação de guerra. Uma condição de parada pode ser quando alguma tropa perder mais de 30% do efetivo.
 - Ex.: como cumprir um contrato de fabricação de 100 aviões em 18 meses?
 - Como o número de observações é finito, a determinação das condições iniciais geralmente afeta a medida de interesse.
- Por outro lado, se não existe um evento que determina a duração da simulação, dizemos que a simulação é “infinita”.
- Se a simulação é infinita e a medida de interesse tem distribuição estacionária, dizemos que a simulação é de “parâmetro estacionário”.

- Ex.: desejamos estimar o número de itens produzidos por hora. A perda de produção é desprezível entre o final do trabalho em um dia e o início no dia seguinte (startup time).
- A maioria dos sistemas reais não têm distribuição estacionária, visto que o sistema muda com o tempo (ex.: número e localização das máquinas).
- Porém a saída de uma simulação (que é uma abstração da realidade) pode ter distribuição estacionária, já que normalmente o modelo assume que as características não mudam com o tempo.

- Considere uma simulação infinita onde a medida não tem distribuição estacionária.

- Suponha que o tempo é dividido em intervalos de igual comprimento, chamados de “ciclos”.
- Seja Y_i^c uma VA definida com base no i -ésimo ciclo.
- Se o processo Y_1^c, Y_2^c, \dots tem distribuição estacionária F^c , dizemos que a simulação é de “parâmetro cíclico”.
- Ex.: no exemplo anterior, se o startup time não for desprezível, a produção por hora não é estacionária. Porém, a produção por dia será.
- Quando a medida tem distribuição estacionária, podemos reduzir o ciclo indefinidamente.
- Como o tempo é discretizada nas análises de parâmetros estacionários, estas análises podem ser empregadas para parâmetros cíclicos. Ex9.29.

- Podemos ter ainda casos onde a simulação é infinita, e nenhum tamanho de ciclo apropriado fornece uma distribuição estacionária.

- Isto acontece quando o modelo muda com o tempo.
- Ex.: o departamento de marketing determina, em função das vendas, um cronograma de 3 meses de produção, onde indica o número e tipos de computadores que devem ser produzidos em cada semana.
- Nestes casos, como normalmente existe uma quantidade finita de dados que descrevem como o modelo muda com o tempo, podemos analisar este simulador como se fosse finito.

50 Análise estatística de simulações finitas

- Neste caso, executamos a simulação até o fim n vezes, e coletamos ao final de cada execução uma medida X .
 - Seja X_i a medida observada na i -ésima execução.

- Como os números aleatórios em cada execução são independentes, e o mesmo estado inicial é mantido, temos que os X_i 's são IID.

50.1 Estimando médias

- Como as VAs são IID, temos os seguintes estimadores não viesados para a média μ e variância σ^2 de X :

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}.$$

Ou seja, $E[\bar{X}] = \mu$ e $E[S^2] = \sigma^2$.

- Pelo Teorema do Limite Central, temos que a distribuição de \bar{X} tende (com o aumento de n) para uma distribuição normal.
- Assim, com confiança $100(1-\alpha)\%$, podemos calcular o seguinte intervalo de confiança para μ :

$$\bar{X} \pm t_{n-1, 1-\alpha/2} \sqrt{\frac{S^2}{n}},$$

onde $t_{n-1, 1-\alpha/2}$ é o ponto crítico com probabilidade acumulada $1-\alpha/2$ em uma distribuição t com $n-1$ graus de liberdade.

- Ex.: 9.14 e 9.15, Law.
- Se a distribuição de X for muito diferente de uma normal, é importante verificar se n é grande o bastante para gerar uma boa aproximação normal para a distribuição da média \bar{X} .

50.1.1 Obtendo determinada precisão

- Para obter maior precisão (tamanho máximo do intervalo de confiança) na estimativa de \bar{X} , basta aumentar o número de execuções n . Como determinar o menor valor de n que garanta uma dada precisão?
- Denotamos $\delta(n, \alpha) = t_{n-1, 1-\alpha/2} \sqrt{S^2/n}$.

Procedimento de amostra fixa

- O “erro absoluto” de \bar{X} é definido como $|\bar{X} - \mu|$.
- Se n é tal que $\delta(n, \alpha) \leq \beta$ (onde $\beta > 0$), então

$$1 - \alpha \approx \Pr(\bar{X} - \delta(n, \alpha) \leq \mu \leq \bar{X} + \delta(n, \alpha))$$

$$= \Pr(|\bar{X} - \mu| \leq \delta(n, \alpha)) \leq \Pr(|\bar{X} - \mu| \leq \beta)$$

- Portanto, para obter um erro absoluto de no máximo β com probabilidade maior que $1-\alpha$, devemos encontrar o menor valor de n tal que $\delta(n, \alpha) \leq \beta$.
- Procedimento:

1. Escolha um valor inicial para n de modo que a estimativa S^2 não mude consideravelmente com o aumento de n .
 2. Utilizando esta estimativa de S^2 , aumente o valor de n até que $\delta(n, \alpha) \leq \beta$.
- A principal dificuldade deste procedimento é identificar um valor inicial apropriado para n no Passo 1, pois a corretude do procedimento depende de S^2 ser uma boa estimativa de σ^2 .
 - Podemos ajustar o procedimento para limitar o “erro relativo” $|\bar{X} - \mu|/|\mu|$. Neste caso, se encontrarmos o menor valor de n tal que $\delta(n, \alpha)/|\bar{X}| \leq \gamma$, temos que

$$\begin{aligned}
 1 - \alpha &\approx \Pr(|\bar{X} - \mu|/|\bar{X}| \leq \delta(n, \alpha)/|\bar{X}|) \\
 &\leq \Pr(|\bar{X} - \mu| \leq \gamma|\bar{X}|) \\
 &= \Pr(|\bar{X} - \mu| \leq \gamma|\bar{X} - \mu + \mu|) \\
 &\leq \Pr(|\bar{X} - \mu| \leq \gamma(|\bar{X} - \mu| + |\mu|)) \\
 &= \Pr\left(\frac{|\bar{X} - \mu|}{|\mu|} \leq \frac{\gamma}{1 - \gamma}\right)
 \end{aligned}$$

- Ou seja, para obter um erro relativo de no máximo γ (com probabilidade de pelo menos $1 - \alpha$), devemos encontrar o menor valor de n tal que $\delta(n, \alpha)/|\bar{X}| \leq \gamma' = \gamma/(\gamma + 1)$.
- Neste caso, o valor inicial para n no Passo 1 deve ser tal que as estimativas \bar{X} e S^2 não modifiquem consideravelmente com o aumento de n .
- Ex.: 9.17 e 9.18, Law.

Procedimento sequencial

- Procedimento:
 1. Escolha um valor inicial $n = n_0$ e execute o simulador n vezes.
 2. Calcule \bar{X} , S^2 e $\delta(n, \alpha)$.
 3. Se $\delta(n, \alpha)/|\bar{X}| \leq \gamma'$, pare. Caso contrário, substitua n por $n + 1$, execute mais uma vez o simulador, e vá para o passo 2.
- Ao contrário do procedimento de amostra fixa, o procedimento sequencial vai atualizando as estimativas \bar{X} e S^2 para cada nova observação (execução do simulador).
- Ex.: 9.19, Law.
- Para maior chance de corretude do procedimento, alguns trabalhos recomendam $n_0 \geq 10$ e $\gamma \leq 0.15$.

Recomendações para escolha dos procedimentos

- Se a precisão do intervalo de confiança não é muito importante, recomenda-se o procedimento de amostra fixa (mais simples).

– Se as observações são muito diferentes de uma normal, e o número de execuções é pequeno, o intervalo obtido pode ser muito diferente do real.

- Caso deseje-se $\gamma \leq 0.15$, o procedimento sequencial é indicado. Caso contrário, recomenda-se aplicações sucessivas do procedimento de amostra fixa.
 - Após estimar n' , colete mais $(n' - n)/2$ observações. Se o intervalo de confiança ainda não possui a precisão desejada, atualize as estimativas \bar{X} e S^2 , e continue o processo.
- As aplicações sucessivas do procedimento de amostra fixa também é recomendado para β pequeno.
- Independente do custo para repetir as execuções, recomenda-se pelo menos de 3 a 5 execuções.

50.2 Escolhendo condições iniciais

- A medida de performance em simulações finitas depende explicitamente das condições iniciais.
- Ex.: a estimativa do tempo de atendimento entre 12h e 13h depende do núm. de clientes no sistema às 12h.
- Uma alternativa é iniciar a simulação em um instante anterior onde as condições iniciais são conhecidas.
 - Ex.: Às 10h nenhum cliente estava no banco.
 - Desvantagem: simular (modelar e executar) período onde os dados não são relevantes.
- Outra alternativa é determinar empiricamente a distribuição de clientes às 12h, e sortear este número para cada execução do simulador.

51 Análise estatística de parâmetros estacionários

51.1 Remoção de transientes iniciais

- Parâmetros estacionários podem ser obtidos com base na distribuição estacionária $F(y)$. Ex.: $E[Y]$.
- Como geralmente não é possível obter as condições que já iniciam o processo em estado estacionário, o estimador do parâmetro com base em um conjunto finito de observações é viesado.
- A abordagem mais sugerida para este problema é a remoção de dados iniciais.
- A técnica mais simples e geral para determinar quantas observações iniciais remover foi proposta por Welch(1981).
- Procedimento (figura 9.5):

1. Execute $n \geq 5$ vezes a simulação, cada execução com m (grande) observações. Seja $Y_{j,i}$ a i -ésima observação da execução j .
2. Seja $\bar{Y}_i = \sum_{j=1}^n Y_{j,i}/n$, $i = 1, \dots, m$. O processo $\bar{Y}_1, \bar{Y}_2, \dots$ tem médias $E[\bar{Y}_i] = E[Y_i]$ e variâncias $\text{Var}[\bar{Y}_i] = \text{Var}[Y_i]/n$. Ou seja, mesma média e $1/n$ da variância original.
3. Caso seja necessário suavizar ainda mais o gráfico de $\bar{Y}_1, \dots, \bar{Y}_m$, geramos o gráfico das médias móveis $\bar{Y}_i(w)$ com janela w .

$$\bar{Y}_i(w) = \begin{cases} \frac{\sum_{s=-w}^w \bar{Y}_{i+s}}{2w+1}, & i = w+1, \dots, m-w \\ \frac{\sum_{s=-(i-1)}^{i-1} \bar{Y}_{i+s}}{2i-1}, & i = 1, \dots, w \end{cases}$$

4. Com base no gráfico, escolha o valor l para o índice i a partir do qual o processo parece ter convergido. Remova os dados com índice menor que l .

- Ex.: 9.25 (Figuras 9.6, 9.7 e 9.8).
- Recomendações para a escolha de n, m, w :
 - Inicialmente faça $n = 5$ ou 10 (dependendo do custo), e m tão grande quanto possível. O valor de m deveria ser grande o bastante para que eventos infrequentes (como quebra de máquinas) ocorram um número razoável de vezes.
 - Gere o gráfico de $\bar{Y}_i(w)$ para vários valores de w , e escolha o menor valor para w de modo que o gráfico fique “razoavelmente suave”.
 - Se nenhum valor de w é satisfatório, gere mais 5 ou 10 execuções de tamanho m . Repita todo o processo até ser capaz de determinar l .

51.2 Método Replicação/Remoção para médias

- Dentre as várias propostas para determinar médias estacionárias, apresentaremos o método R/R, pois:
 - É o mais fácil de entender e implementar.
 - Fornece boa performance estatística.
 - Pode ser aplicado para todos os tipos de parâmetros, e permite estimar vários parâmetros de uma mesma simulação.
- Procedimento:
 1. Determine a quantidade mínima de observações iniciais l que devem ser removidas (Welch).
 2. Produza novas observações Y_{ji} executando a simulação n' vezes, coletando m' observações por execução. m' deve ser muito maior que l .
 3. Seja

$$X_j = \frac{\sum_{i=l+1}^{m'} Y_{ji}}{m' - l}, \quad j = 1, \dots, n'.$$

4. Como as observações com índice maior que l são consideradas convergentes, $E[X_j] \approx E[Y]$. Utilizando os X_j 's, estimamos \bar{X} , S^2 e o intervalo de confiança para Y :

$$\bar{X} \pm t_{n'-1, 1-\alpha/2} \sqrt{\frac{S^2}{n'}}.$$

- Se m for muito maior que l , podemos utilizar o mesmo conjunto de observações para determinar l e o intervalo de confiança. Porém, utilizar conjuntos independentes é mais correto estatisticamente.
- Ex.: 9.27.

52 Estimando proporções

- Em alguns casos, não conseguimos tomar a decisão com base apenas nas médias.
- Ex.: No exemplo 9.20, dois sistemas apresentaram as mesmas médias (tabela 9.4). Entretanto, o histograma (tabela 9.5) revelou que o sistema com 5 filas tem maior dispersão (com mais clientes aguardando muito tempo).
- Como observado no exemplo, o uso de proporções é uma alternativa ao uso de médias.

- Seja

$$Y_i = \begin{cases} 1, & X_i \in B \\ 0, & \text{caso contrário} \end{cases}$$

- Então, $S = \sum_{i=1}^n Y_i$ é o número de observações X_i que pertencem ao conjunto B .
- Se as observações são IID, S tem distribuição binomial com parâmetros n e $p = \Pr(X \in B)$.
- Desejamos entrar um intervalo de confiança para $\hat{p} = S/n$. Note que $E[S/n] = p$.
- Para n grande, S tem distribuição aproximadamente normal com média np e variância $np(1-p)$. Portanto,

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

tem distribuição aprox. normal padrão.

- Portanto, temos o seguinte intervalo de confiança para \hat{p} ,

$$\hat{p} \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}},$$

onde podemos substituir p por \hat{p} .

53 Múltiplas medidas de performance

blank

- Geralmente estamos interessados em coletar várias medidas de performance de uma mesma simulação.
- Suponha que geramos (com base em execuções do simulador), para cada medida μ_s , um intervalo de confiança I_s com confiança $100(1 - \alpha_s)\%$. Qual a confiança de que todas as medidas estão contidas em seus respectivos intervalos de confiança?
- Seja A_s o evento que indica que a medida μ_s **não** pertence a I_s . Então, de acordo a desigualdade de Boole,

$$\Pr\left(\bigcup_{s=1}^n A_s\right) \leq \sum_{s=1}^n \Pr(A_s) = \sum_{s=1}^n \alpha_s.$$

- Portanto,

$$\Pr(\mu_s \in I_s, \text{ para todo } s = 1, \dots, n) \geq 1 - \sum_{s=1}^n \alpha_s.$$

- Este resultado não é animador para o caso de muitas medidas. Ex.: 10 medidas, todas com $\alpha = 0.1$. Assim, só podemos garantir que a probabilidade (de todas pertencerem aos intervalos de confiança) é maior ou igual a zero.
- Se n não é grande, podemos escolher os α_s de modo que $\sum_{s=1}^n \alpha_s = \alpha$, onde α é a probabilidade conjunta desejada.
 - Podemos utilizar valores menores de α_s para medidas μ_s mais importantes.
 - A desvantagem é que os I_s serão maiores, ou mais dados serão necessários.
 - Recomenda-se que $n \leq 10$.
 - Ex.: 9.30 (figura 9.9).