



מרצה : פרופסור אבישי מנדלבאום

מתרגלת : ניצן כרמלי

תאריך הבחינה : 13.2.2014

מס' סטודנט _____

הנדסת מערכות שירות – 096324

מועד א' – סמסטר חורף תשע"ד 2014

פתרון

- המבחן נמשך שלוש שעות. באחריותך להחזירו במועד.
- המבחן בחומר סגור.
- מותר להשתמש במחשבון.
- המבחן כולל 32 עמודים, כולל עמוד זה ונוסחאות מהסתברות בעמוד האחרון.
- כאשר נדרשים חישובים, ניתן לעגלם באופן שאינו פוגע בהבנת התשובה.
- מלאו את הפרטים הנדרשים בראש העמוד, בכתב ברור בבקשה.
- על התשובות יש לענות במקומות המיועדים, שאמורים להספיק: מלל מיותר יגרע ניקוד.
- יש לתת הסברים או הוכחות רק אם התבקשתם במפורש לעשות זאת.

הרצאות/תרגולים/בחינות קודמות	מתוך 21 נקודות	שאלה 1	ניקוד :
תרגילי בית	מתוך 21 נקודות	שאלה 2	
יישום	מתוך 42 נקודות	שאלה 3	
תיאוריה	מתוך 16 נקודות	שאלה 4	

סה"כ _____ מתוך 100 נקודות

בהצלחה !

שאלה 1 - תרגילי כיתה/הרצאות/מבחנים קודמים (21 נקודות)

חלק 1. מופעים למערכת שירות ותחזיות

1.1.1 הגרפים בעמוד הבא מתארים מופעים למוקד שירות טלפוני בארץ, בימים שונים של השבוע: גרף 1 מתאר מספר מופעים לחצי שעה, גרף 2 מתאר את מספר מופעים זה ביחס לממוצע היומי (באחוזים). בהנחה שהמופע למוקד הוא פואסון לא הומוגני בזמן, באיזה מודל מתמטי לקצב המופע, $\lambda(t)$, הייתם ממליצים להשתמש, ובאלו שיטות הייתם חוזים אותו?

תשובה:

אנו ממליצים על שימוש במודל $\lambda(t) = C\lambda_0(t)$ כאשר המ"מ C הינו מספר המופעים ביום מסוים והפונקציה $\lambda_0(t)$ היא הצורה של המופעים, המשתנה עבור ימי חול, ימי שבת וימי ראשון. אפשר לחזות את C ע"פ שיטות פשוטות כגון שיטת הממוצע הנע:

$$F_{j,k} = \frac{N_{j-7,k} + N_{j-14,k} + N_{j-21,k} + N_{j-28,k}}{4}$$

או $F_{j,k} = N_{j-7,k}$: Most Recent Observation

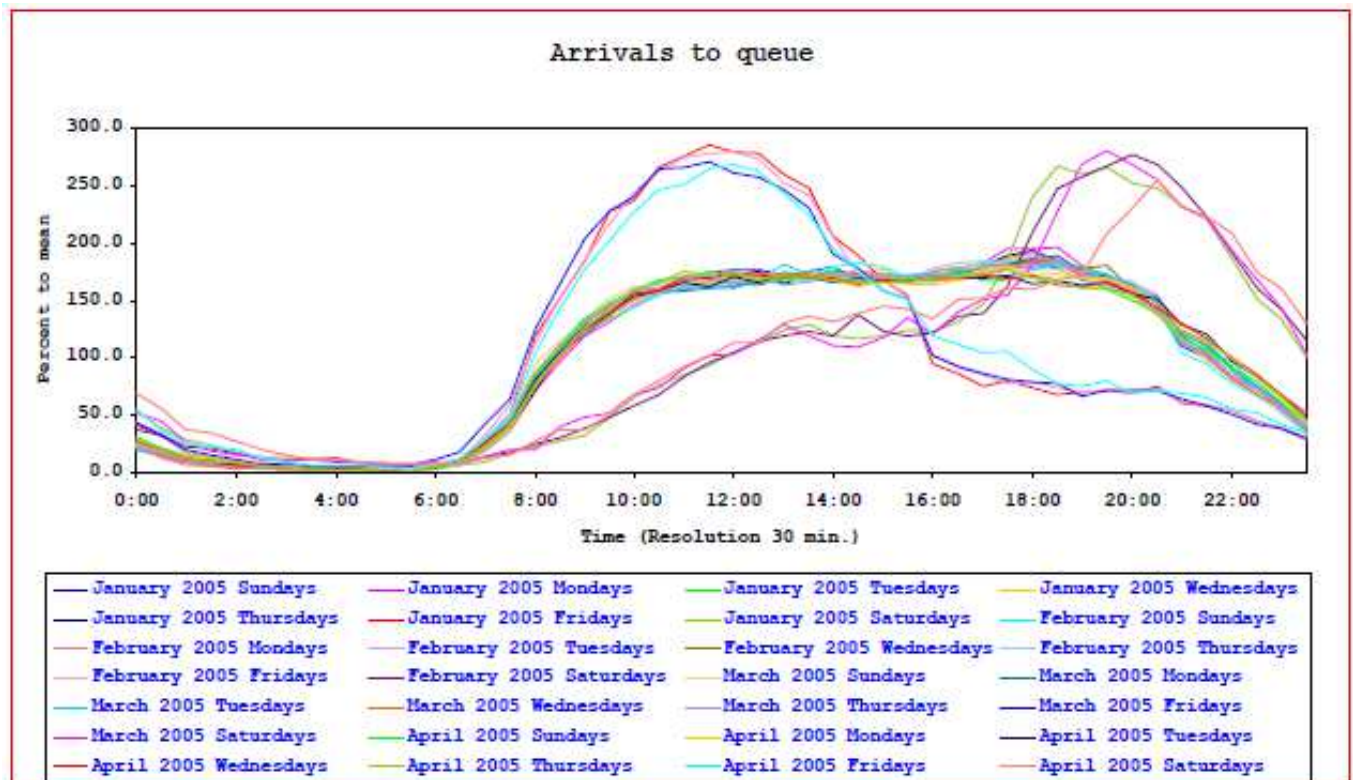
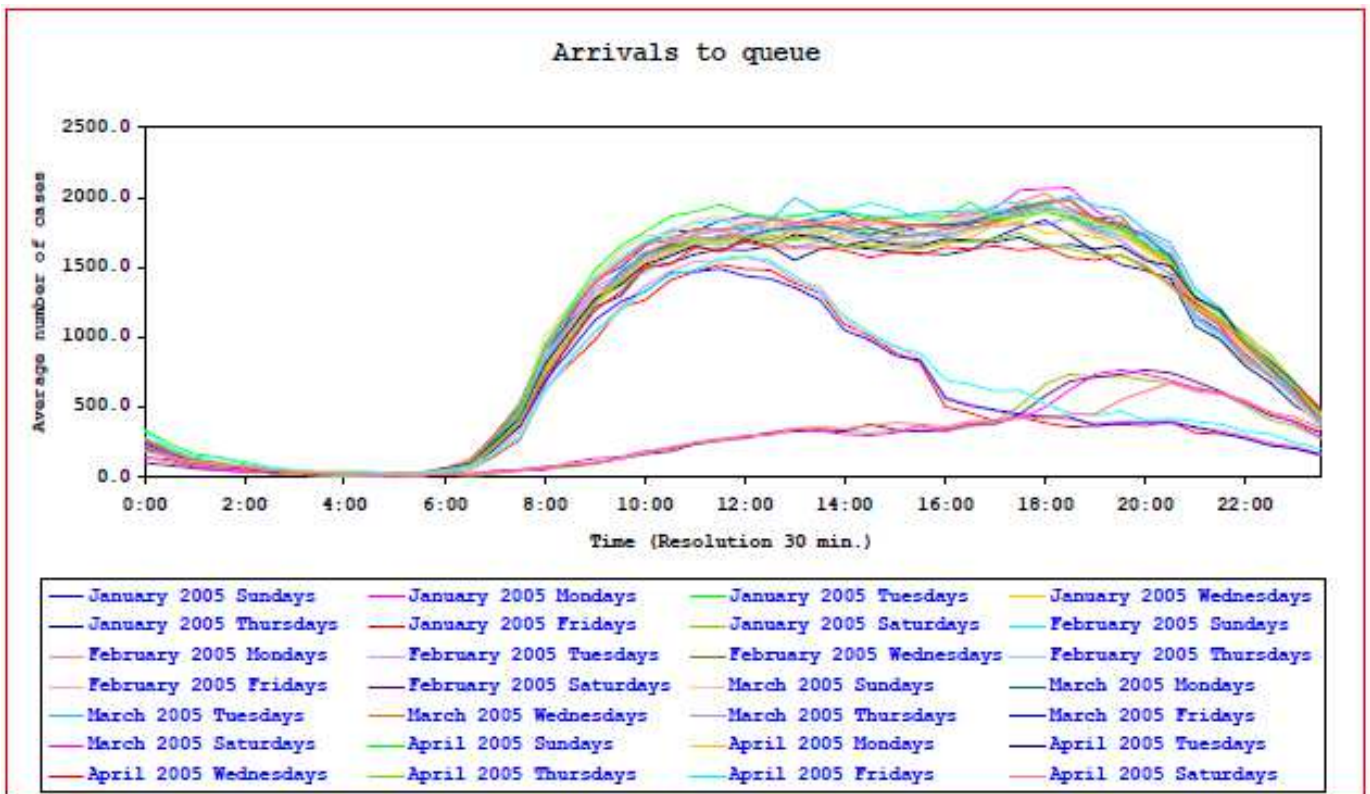
או ע"פ שיטות מתחום סדרות עתיות.

על פי "צורת" היום המתקבלת בגרף השני, ניתן למצוא את $\lambda_0(t)$, בהתאם ליום המסוים בשבוע.

1.1.2 בגרף 3 (בעמוד 5) מתוארים שלושה ניסיונות לחזות מופע למוקד שירות טלפוני בארה"ב. התחזית היא למספר המופעים החל משעה 12:00. הנקודות מתארות את המספר האמיתי. סכמו בקצרה את המסקנות העיקריות הנובעות מדוגמה זו.

תשובה:

נתונים מהבוקר (7:00-10:00) מאפשרים לשפר את איכות החיזוי בצורה משמעותית (ביחס לתחזיות ע"פ נתונים עד יום קודם). אין הבדל משמעותי בין תחזיות על סמך נתונים עד 10 בבוקר ותחזיות על סמך נתונים עד 12 בצהריים. ז"א, השעות 10:00-12:00 לא מוסיפות אינפורמציה.



1.1.3 בגרף 4 בעמוד הבא מתואר קצב המופע של לקוחות שפנו למוקד שירות טלפוני במהלך יום מסוים. **נקודות** בגרף מתארות את מספרי המופעים על פני אינטרוולים של חמש דקות. **תחזית** למספרי מופעים אלה שחושבה על סמך נתונים מימים קודמים מתוארת ע"י **הקו**. נסמן ב- K את מספר האינטרוולים של חמש דקות במהלך יום עבודה. נסמן ב- N_k את מספר המופעים בפועל על פני אינטרוול k , $1 \leq k \leq K$, ונניח ש- N_k מתפלג לפי התפלגות פואסון עם פרמטר לא ידוע λ_k .

בגרף אנו גם רואים שני סוגים של רווחי סמך. אנו נשתמש בשם **רווח סמך צר** עבור רווח הסמך הצר יותר (אפור בהיר) ובשם **רווח סמך רחב** עבור רווח הסמך הרחב יותר (אפור כהה). אחד מרווחי הסמך האלה מתאים לקצבי המופע λ_k והשני מתאים למספרי המופעים בפועל N_k . איזה רווח סמך מתאים ל- λ_k ואיזה ל- N_k ? נמקו את תשובתכם בקצרה.

תשובה:

רווח הסמך הצר מתאים ל- λ_k ורווח הסמך הרחב מתאים ל- N_k .
 $N_k \sim \text{Poisson}(\lambda_k)$ ולכן ההשתנות של λ_k קטנה יותר. באמידת N_k עלינו לאמוד גם את הפרמטר λ_k וגם את הריאליזציה לכמות המופעים בפועל בכל אינטרוול.

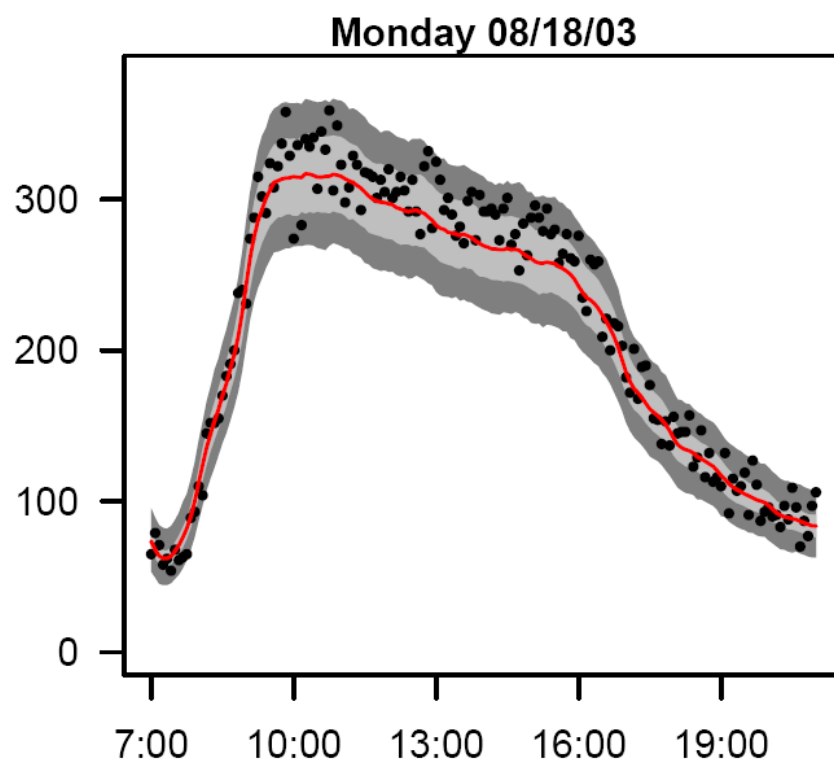
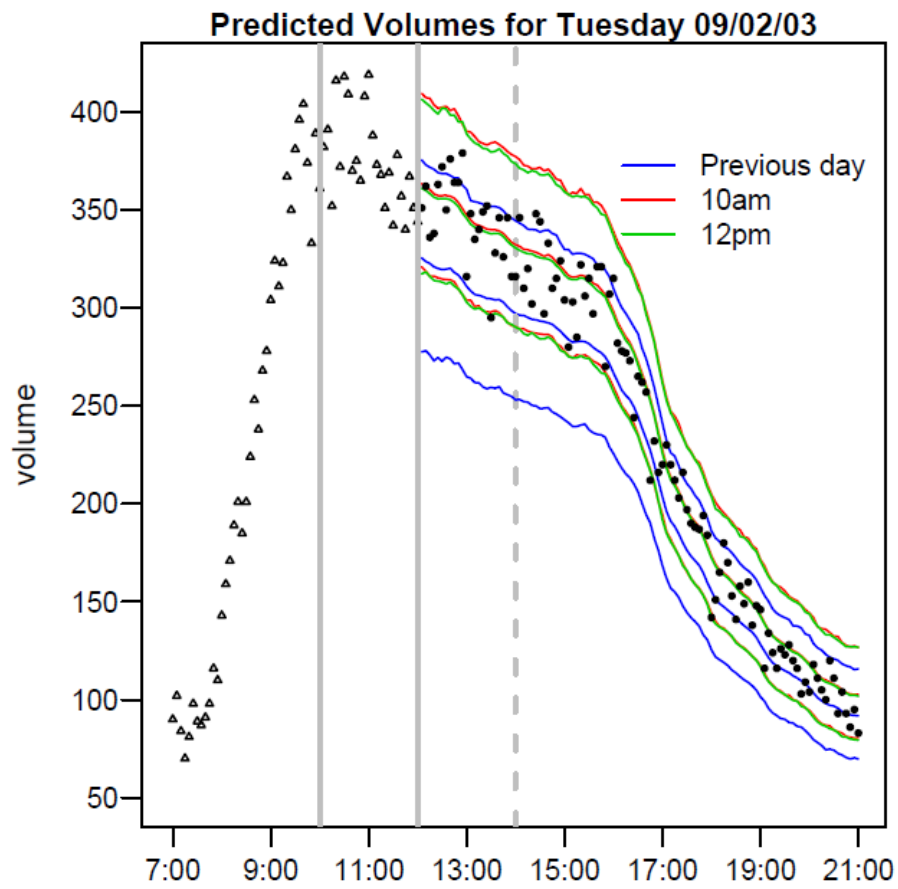
1.1.4 נניח שהפרמטרים λ_k ידועים בוודאות. איך ישתנו במקרה זה רווחי הסמך? סמנו ע"י עגול את התשובה הנכונה. הסבירו בקצרה את תשובתכם.

תשובה:

- רווחי הסמך לא ישתנו.
- רווחי הסמך הצרים והרחבים יהיו זהים זה לזה.
- רווח הסמך הצר יעלם ורווח הסמך הרחב יהיה צר יותר.
- רווח הסמך הצר יעלם ורווח הסמך הרחב לא ישתנה.

הסבר:

אם הפרמטרים λ_k ידועים בוודאות אזי רווח הסמך לפרמטר הוא פשוט הערך של הפרמטר עצמו. רווח הסמך עבור N_k יהיה כעת צר יותר כי אין צורך לאמוד את הפרמטר של ההתפלגות, רק את הריאליזציה.



חלק 2. פונקציית סיכון

1.2.1 עבור מ"מ X אי שלילי עם פונקציית התפלגות F_X ופונקציית צפיפות f_X , הגדירו את פונקציית הסיכון h_X . הסבירו בקיצור ובמדויק את משמעות $h_X(t)$, בזמן t כלשהו.

תשובה:

$$h_X(t) = \frac{f_X(t)}{1 - F_X(t)}, \quad t \geq 0$$

הסבר: $h_X(t) \cdot \Delta \approx P\{X \leq t + \Delta \mid X > t\}$. כלומר, פרופורציונאלי לסיכון הרגעי לחוות אירוע מיד אחרי זמן t בהינתן שעד זמן t לא התרחש אירוע.

1.2.2 נתון מ"מ X מעריכי עם פרמטר λ : $X \sim \exp(\lambda)$. צפיפותו נתונה ע"י $f_X(t) = \lambda e^{-\lambda t}$, $t \geq 0$. חשבו את פונקציית הסיכון של X . תנו הסבר אינטואיטיבי קצר לתשובתכם.

הוכחה: $h_X(t) = \frac{f_X(t)}{P(X > t)} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda$. ז"א פונקציית הסיכון קבועה בזמן.

הסבר: תכונת "חוסר הזיכרון" של משתנה מקרי מעריכי.

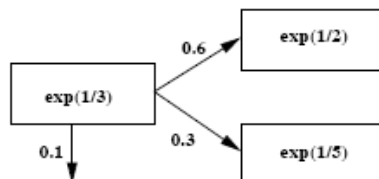
1.2.3 בעמוד הבא מצורף דף מהתרגול. הסבירו מדוע הגרף מתחיל בערך 0.033.

תשובה:

השרות יסתיים מיד רק אם השרות יסתיים לאחר השלב הראשון (בהסתברות 0.1) וגם המשתנה המקרי המתאר את משך השלב הראשון יקבל ערך קרוב ל-0. זה יקרה בהסתברות: $1/3 \cdot 0.1$.

Theoretical Calculation and Statistical Estimation of Hazard Rate

Example: consider the following service time distribution:



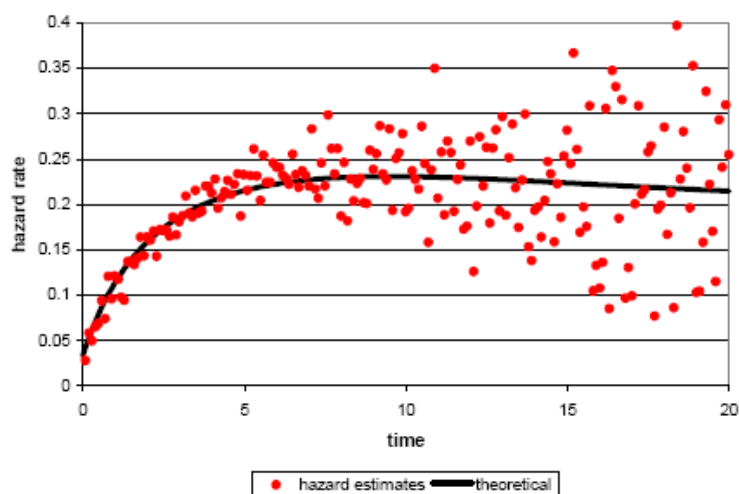
Its hazard rate can be calculated theoretically:

$$h(x) = \frac{0.15 \cdot e^{-x/5} + (29/60) \cdot e^{-x/3} - 0.6 \cdot e^{-x/2}}{0.75 \cdot e^{-x/5} + 1.45 \cdot e^{-x/3} - 1.2 \cdot e^{-x/2}}.$$

How do we estimate hazard rate from data?

Description of simulation experiment.

10,000 independent realizations of service time above were simulated in Excel. The theoretical hazard rates were plotted and compared against estimates of the hazard rate, based on the simulation data. (The method used for hazard rate estimation is described on the next page.)



1.2.4 לאיזה ערך שואף הגרף כאשר $t \rightarrow \infty$? הסבירו אינטואיטיבית תשובתכם.

תשובה:

הגרף שואף לערך 0.2. בהינתן ששרדנו זמן רב, הסבירות היא גבוהה להיות בשלב השני הארוך יותר. התפלגות השלב השני הארוך היא $\exp(1/5)$, ופונקציית הסיכון של משתנה מקרי מעריכי היא קבועה ושווה לפרמטר ההתפלגות λ , כפי שראינו בסעיף 1.2.2

שאלה 2 – תרגילי בית (21 נקודות)

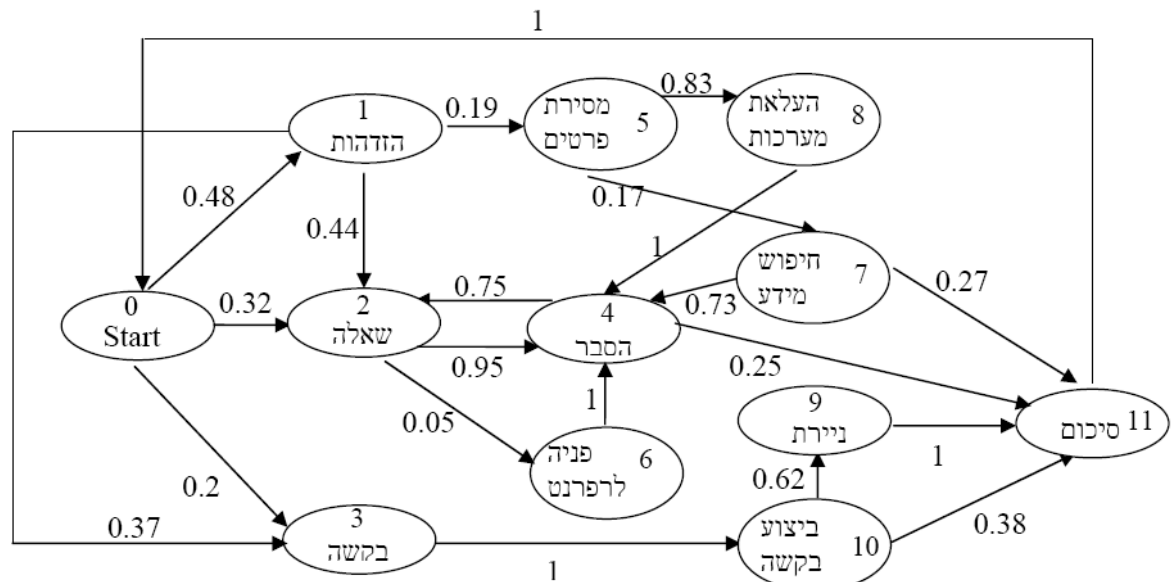
חלק 1 - התפלגות זמני שהיה

בשנת 2007 בוצע בפקולטה לתעשייה וניהול פרויקט שנתי בשיתוף פעולה עם מוקד טלפוני של בנק ישראלי גדול. אחת ממטרות הפרויקט הייתה: חקר תהליך ביצוע שיחות מסוג שירות מסוים. על מנת לאפיין את תהליך השיחה ביחידת שירות זו, בוצעו מדידות ראשוניות על ידי האזנה למוקדנים. לאחר ניתוח השיחות, נמצא כי ניתן לחלק את השיחה לפאזות עיקריות (אלמנטים). אלמנטים אלו מתאימים לכל השיחות בשירות הנדון.

האלמנטים העיקריים המרכיבים שיחה (אין צורך להתעמק בתיאור האלמנטים):

1. **הזדהות**- אם ללקוח יש קוד סודי והוא מקיש אותו במענה הקולי אזי הזיהוי הוא אוטומאטי. אם הקוד לא הוקש, מתבקש הלקוח להזדהות מול המוקדן ולענות על מספר שאלות לאימות הנתונים.
2. **שאלה**- שאלה של לקוח. לדוגמא הבהרה של מכתב שהתקבל מהבנק, בירורים לגבי אופן התשלום, ריבית ועוד.
3. **בקשה**- בקשה של הלקוח.
4. **הסבר**- מתן הסבר על-ידי הבנקאי לשאלות הלקוח.
5. **מסירת פרטים**- מסירת פרטים על-ידי הלקוח לבנקאי, כדי לאפשר לבנקאי לבצע פעולות מסוימות, לדוגמא משלוח פקס עם פרטי הלקוח.
6. **פנייה לרפרנט**- אם הבנקאי לא מצא תשובה, יש באפשרותו לפנות לבנקאי ותיק יותר (הרפרנט) ולבקש ממנו עזרה. הפנייה לרפרנט **קוטעת** את השיחה.
7. **חיפוש מידע**- מעבר על היסטוריה של הלקוח לצורך מציאת מידע.
8. **העלאת מערכות**- בנקאים עובדים מול שלוש מערכות. על מנת להתחיל את תהליך הטיפול בלקוח, יש צורך לחכות ששלושת המערכות תעלנה על מסך הבנקאי. זמן המתנה זה כלול באלמנט.
9. **ניירת**- כתיבת הוראות במחשב כגון- הוראה למשלוח פקס ועוד. כמעט ואין ביצוע ניירת פיזית כגון שליחת מכתבים.
10. **בצוע בקשה**- ביצוע בקשת הלקוח על-ידי בנקאי, כוללת את כל הפעולות הנדרשות מהבנקאי, לדוגמא שליחת מכתב, ביצוע שינויים ועוד.
11. **סיכום**- תיעוד השיחה במחשב.

בעזרת נתוני ההאזנות, חושבו פרופורציות המעברים (הסתברויות מעבר) בין האלמנטים (המצבים). שרשרת מרקוב, כמתואר להלן, מהווה מודל לתהליך ביצוע האלמנטים על ידי מוקדן כלשהו, ממחלקת השירות שנחקרה בפרויקט.



נניח שההתפלגות הסטציונרית של השרשרת נתונה על ידי $\pi = (\pi_0, \pi_1, \dots, \pi_{11})$. לאחר הגדרת האלמנטים (הפאזות) בשיחה, בוצע חקר זמנים לצורך אמידת ממוצע משכי האלמנטים והתפלגותם.

עבור כל אלמנט התקבלו הממוצעים הבאים:

מספר	תיאור אלמנט	ממוצע משך האלמנט (שניות)
1	הזדהות	46
2	שאלה	120
3	בקשה	44
4	הסבר	180
5	מסירת פרטים	64
6	פניה לרפרנט	260
7	חיפוש מידע	20
8	העלאת מערכות	46
9	ניירת	61
10	ביצוע בקשה	20
11	סיכום	10

לצורך פתרון השאלות הבאות, ניתן להניח שהתפלגויות משך כל אלמנט הן מעריכיות עם תוחלות הנתונות בטבלה (המציאות לא רחוקה מהנחה זו). בנוסף, התברר ששיחות למוקדן מגיעות לפי תהליך פואסון, בקצב של 12 שיחות לשעה. אם מוקדן עסוק, השיחות עוברות למוקדן אחר.

לאחר חקר הזמנים, משתתפי הפרויקט הגיעו למסקנה שאפשר לתאר את תהליך השיחות שמבצע מוקדן על-ידי תהליך קפיצה מרקובי $X = \{X_t, t \geq 0\}$, כאשר X_t מתאר את המצב (אלמנט, פאזה) של שיחה בזמן t .

2.1.1 עבור התק"מ X , נסמן ב- P את מטריצה הסתברויות המעבר שלו וב- q את וקטור קצבי עזיבה. השלימו את הערכים הבאים:

$p_{12} = 0.44$	$q_0 = 12$
$p_{11} = 0$	$q_2 = \frac{3600}{120} = 30$
$p_{26} = 0.05$	$q_4 = \frac{3600}{180} = 20$
$p_{24} = 0.95$	$q_{11} = \frac{3600}{10} = 360$

2.1.2 כיצד מתפלג משך שיחה? תארו במדויק ככל שתוכלו.

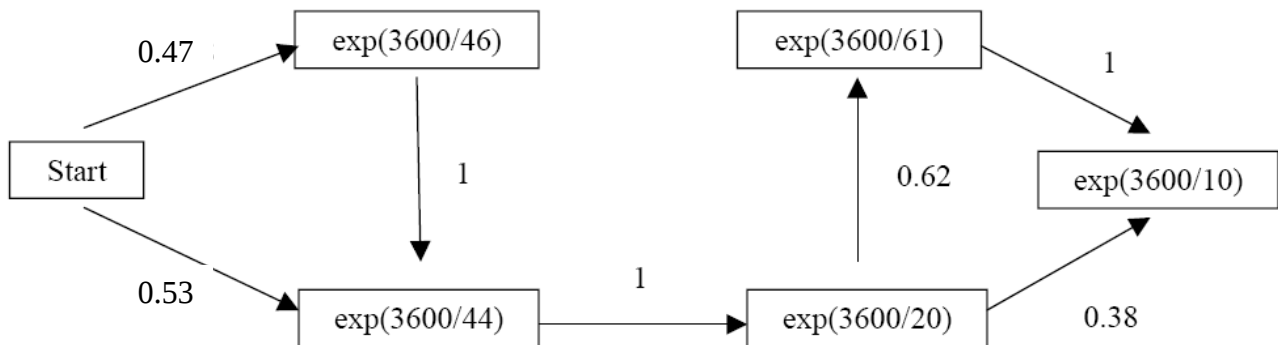
תשובה:

משך השיחה מתפלג לפי התפלגות Phase Type כך שהמעברים בין הפאזות מוגדרים בדיאגרמה שנתונה בתחילת השאלה, וזאת ללא מעבר ממצב 11 למצב 0. זמני השהייה בכל הפאזות הם מעריכיים עם התוחלות הנתונות בטבלה.

2.1.3 מהי תוחלת משך השיחה של לקוח שמתקשר עם "בקשה"?

תשובה:

אחוז השיחות שעוברות בקשה מכלל השיחות הוא: $0.2 + 0.48 \cdot 0.37 = 0.3776$, כאשר 53 אחוז מהן עוברות ישירות ל"בקשה" $\left(\frac{0.2}{0.3776} = 0.53\right)$ ו-47% דרך הזדהות $\left(\frac{0.37 \cdot 0.48}{0.3776} = 0.47\right)$.



נגדיר T - זמן שיחה עם "בקשה".

$$E[T] = 44 + 0.47 \cdot 46 + 20 + 0.62 \cdot (61 + 10) + 0.38 \cdot 10 = 133.44$$

2.1.4 אם מוקדן נמצא עכשיו בשלב של "העלאת מערכות" (אלמנט 8), מהי ההסתברות שלא יבקר באלמנט 6 ("פניה לרפרנט") עד סוף השיחה?

תשובה:

$$\begin{aligned}
 P(\dots) &= 0.25 \cdot (1 + 0.75 \cdot 0.95 + (0.75 \cdot 0.95)^2 + (0.75 \cdot 0.95)^3 + \dots) = \\
 &= 0.25 \cdot \frac{1}{1 - 0.75 \cdot 0.95} = 0.869565
 \end{aligned}$$

חלק 2. מודל נוזלים

Tele-SHOP היא חברה העוסקת במכירות המתבצעות באמצעות מוקד שירות טלפוני. ניתן לתאר את הגעת השיחות למוקד השירות ע"י תהליך פואסון לא הומוגני בזמן עם פונקציה הקצב $\lambda(t)$.

המוקד פועל בשעות 16:00-24:00. נסמן אינטרוול זמן זה ב- $[0, T]$.

ממוצע משך שירות של שיחה הינו $\frac{1}{\mu}$ דקות וממוצע סבלנותו של לקוח (עד נטישה) הינו $\frac{1}{\theta}$.

הניחו כי עלות העסקת שרת הינה c ש"לדקה ועלות המתנה הינה h ש"לדקה. עוד הניחו כי הרווח מסיום שירות הינו r ש"לדקה והקנס על נטישת לקוח הינו s ש"לדקה.

נסמן ב- $Q(t)$ את מספר הלקוחות במערכת בזמן $0 \leq t \leq T$,

וב- $N(t)$ את מספר המוקדנים במערכת בזמן $0 \leq t \leq T$.

2.2.1 כתבו משוואה דיפרנציאלית עבור $Q(t)$, $0 \leq t \leq T$.

תשובה:

$$\frac{dQ(t)}{dt} = \lambda(t) - \mu[Q(t) \wedge N(t)] - \theta[Q(t) - N(t)]^+, \quad t \geq 0$$

2.2.2 תארו בקיצור איך ניתן, בעזרת גיליון אלקטרוני בלבד, לפתור בקירוב (מדויק ככל הנדרש) את המשוואה הדיפרנציאלית בסעיף 2.2.1.

תשובה:

נתחיל עם $Q(0)$ נתון. לאחר מכן, עבור $t_n = t_{n-1} + \Delta$, $n = 1, 2, \dots$, נחשב,

$$Q(t_n) = Q(t_{n-1}) + \lambda(t_{n-1})\Delta - \mu[Q(t_{n-1}) \wedge N(t_{n-1})]\Delta - \theta[Q(t_{n-1}) - N(t_{n-1})]^+ \Delta$$

ונקבל את $Q(0), Q(\Delta), Q(2\Delta), \dots$. ככל ש- Δ קטן יותר, כך הקירוב מדויק יותר.

2.2.3 רשמו ביטוי לאחוז הנוטשים באינטרוול $[0, T]$.

תשובה:

$$P\{Ab\} = \frac{\int_0^T \theta(Q(t) - N(t))^+ dt}{\int_0^T \lambda(t) dt}$$

2.2.4 רשמו ביטוי לרווח הכולל במערכת עד זמן T .

תשובה:

נסמן:

c – עלות מוקדן ליח' זמן, h – עלות המתנת לקוח ליח' זמן, r – רווח מסיום שירות, s – קנס על נטישת לקוח.

$$C(N) = \int_0^T \{ r\mu [Q(t) \wedge N(t)] - (s\theta + h)[Q(t) - N(t)]^+ - cN(t) \} dt$$

הניחו כעת כי מספר השרתים קבוע במהלך יום העבודה. בתרגיל הבית, בעזרת Excel solver, מצאתם $N^* =$ מספר השרתים האופטימלי שמביא למקסימום את רווח המוקד.

2.2.5 כיצד ישתנה הערך N^* כאשר ערכו של הרווח מסיום שירות בודד, r , עולה? תארו, בדיוק רב ככל שתוכלו, את התנהגות N^* כאשר r גדל לאינסוף.

תשובה:

כאשר r גדל ועלויות המערכת נותרות קבועות, כדאי להעסיק יותר שרתים כדי למנוע נטיות מהמערכת.

כאשר r גדל לאינסוף, N^* מתכנס לערך סופי, כאשר

$N^* = \max R(t)$ ו- $R(t)$ הוא פתרון המשוואה הדיפרנציאלית $\frac{dR(t)}{dt} = \lambda(t) - \mu R(t)$, כלומר

$R(t)$ הוא מספר הלקוחות במערכת עם אינסוף מוקדנים = העומס המוצע (Offered Load) על המערכת בזמן t .

שאלה 3 – יישום (42 נקודות)

חוק הקנאביס הרפואי

להלן כתבה שהתפרסמה ב-YNET ב-15.12.13, **אין צורך לקרוא את הכתבה**, כל המידע הדרוש יופיע בשאלות עצמן.

סוף לתורים? הממשלה אישרה חוק הקנאביס הרפואי

על פי הנוהל החדש שאושר יתווספו עשרה רופאים שיוכלו לתת היתר לשימוש בקנאביס רפואי, מה שצפוי לקצר את זמן ההמתנה. עוד הוחלט: הקנאביס יינתן בבתי מרקחת, חולה סופני יקבל אישור תוך 48 שעות וחולי סרטן העוברים טיפולי הקרנה וכימותרפיה - בתוך שבוע.

Recommend 802

ירחן קלנר פורסם: 15.12.13, 19:41

האם הסאגה סביב נושא הקנאביס הרפואי בישראל בדרך לסיום? הממשלה אישרה היום (א') הצעת מחליטים שגיבשו משרדי הבריאות, ביטחון הפנים והחקלאות. לפי ההחלטה, [שעיקרה פורסמו ב-ynet](#), ייקבעו זמנים מחייבים למתן שירות, עשרה רופאים נוספים יתווספו לרשימת המוסמכים לאשר רישיון לשימוש בקנאביס רפואי וחברת "שרל", הנמצאת בבעלות בתי החולים. תיכנס כגורם נוסף בין מגדלי הקנאביס לבין החולים.

מה כולל החוק החדש?

- **הקנאביס יסופק בבתי מרקחת - ולא אצל המגדלים:** במסגרת תהליך ההסדרה חברת "שרל" תאסוף את הקנאביס מהמגדלים ותעביר אותו למרכז לוגיסטי, שם ייארז במינונים קבועים לפי סטנדרטים שיקבע משרד הבריאות. משם הקנאביס הרפואי יעבור לבתי מרקחת ייעודיים.
- **תוספת של 10 רופאים:** עד היום היו תורים ארוכים למתן אישור קנאביס רפואי בשל מיעוט הרופאים. כעת צופים שהתוספת תגרום לקיצור התורים, אך המבקרים טוענים כי מספר הרופאים שיתווספו קטן - וההשפעה שלו תהיה שולית.
- **לוח זמנים מחייב:** חולה סופני יקבל אישור לקנאביס רפואי בתוך 48 שעות בלבד. חולה סרטן המקבל כימותרפיה או הקרנות - עד שבוע. הודעה על כך שהבקשה למתן קנאביס התקבלה אצל הרופא תינתן בתוך שני ימי עבודה, ובהמשך ייקבע תוך כמה זמן תהיה חובה על מתן תשובה לחולה.

אם בשנת 2003 קיבלו 60 חולים אישורים לשימוש בקנאביס רפואי, הרי שכעבור שלוש שנים הגיע מספרם ל-200-300, כשבמקביל החלו לפעול מגדלים מקומיים; בשנת 2010 עלה מספרם ל-2,000-3,000; ובשנה האחרונה מדובר כבר ב-11 אלף איש שמקבלים את האישור

כבר עשר שנים שחולים ישראלים יכולים להשתמש בקנאביס למען מטרות רפואיות, אבל הוויכוח על הסדרת הנושא עדיין בעיצומו. הדרישה לקבל את הקנאביס הרפואי היא גבוהה ועולה בקביעות, ומנגד מנסה משרד הבריאות מנסה לספק הקלה כלשהי לחולים הקשים, תוך ניסיון להסדיר את התחום החדש.

בשנת 2010, ע"פ נתוני הכתבה, כ – 3000 חולים קיבלו אישור לצריכת קנאביס רפואי. בשנת 2013, 11,000 חולים קיבלו אישור שכזה. נניח כי ב – 2010 אחוז הבקשות המאושרות מתוך כלל הבקשות שנכנסו באותה שנה היה 75% ובשנת 2013 רק שני שליש מתוך כלל הבקשות שנכנסו אושרו והשאר נדחו. בנוסף נניח כי הזמן הדרוש לרופא לטפל בבקשה לא השתנה מאז שנת 2010.

3.1 מה היה הגידול באחוזים בעומס העבודה (העומס המוצע) משנת 2010 ועד שנת 2013 הנובע מבקשות לאישור שימוש בקנאביס רפואי?

תשובה:

בשנת 2006 קצב כניסת בקשות למערכת היה: $\lambda = \frac{3000}{0.75} = 4,000 \text{ requests / year}$

בשנת 2013 קצב כניסת בקשות למערכת היה: $\lambda = \frac{11,000}{(2/3)} = 16,500 \text{ requests / year}$

תוחלת משך השירות לא השתנתה ולפיכך:

$$\frac{R_{2013}}{R_{2006}} = \frac{16,500 \cdot E(S)}{4,000 \cdot E(S)} \times 100 = 412.5\%$$

עליה של מעל 400 אחוזים.

3.2 בשנת 2013, נניח כי משך השהייה של בקשה במערכת, מרגע פתיחתה ועד קבלת החלטה בעניינה הוא 4 חודשים. מהו ממוצע מספר הבקשות הפתוחות במערכת בשנה זו?

תשובה:

מצאנו בסעיף הקודם כי $\lambda = 16,500 \text{ requests / year}$

נתון $E(W) = 4 \text{ months} = 1/3 \text{ years}$

$$E(L) = \lambda E(W) = \frac{16,500}{3} = 5,500 \text{ requests}$$

מחוק Little:

תהליך קבלת האישור לשימוש בקנאביס רפואי כולל מספר שלבים. לצורך השאלה, נפשט את התהליך באופן הבא: כאשר חולה מקבל מרופאו האישי המלצה לשימוש בקנאביס רפואי, הבקשה מועברת לאחד מ-21 רופאים המוסמכים ע"י משרד הבריאות לאשר שימוש בקנאביס רפואי. בשאלה זו נתמקד במערכת של 21 הרופאים המוסמכים למתן אישורים מסוג זה. נציין כי רופאים אלו מטפלים בנושא האישורים, בנוסף לתפקידים נוספים שהם מבצעים במהלך שגרת העבודה העמוסה שלהם. עובדה זו יוצרת תורים ארוכים. כחלק מהניסיון של משרד הבריאות לקצר תורים, נקבעה תוספת של 10 רופאים לסך הרופאים המוסמכים למתן אישור שכזה. **תוספת זו נגזרה מתוך תוספת התקציב שאושרה לכך.** במקביל ובאופן בלתי תלוי, נקבעו אילוצי זמנים למתן אישור בהתאם לסוג החולה. בשאלה זו ננסה לבחון באמצעות כלים שנלמדו בקורס, מהי תוספת הרופאים שאכן נדרשת כדי לעמוד באילוצים אלו.

3.3 נניח כי, על פי התחזיות, כ – 20,000 בקשות לשימוש בקנאביס רפואי עתידות להיכנס למערכת בשנת 2014, ובקשות אלו תשארנה בתור עד לטיפולן. כל רופא יכול לטפל ב-2 בקשות ביום (40 בקשות בחודש, 480 בקשות בשנה). תחת ההנחות עד כה (בפרט עם 21 רופאים), האם המערכת תהיה יציבה? האם המערכת תהיה יציבה דרוש: $\rho < 1$

תשובה:

$$\rho = \frac{\lambda}{n\mu} = \frac{20,000}{21 \cdot 480} = 1.984 > 1$$

על מנת שהמערכת תישאר יציבה דרוש: $\rho < 1$

$$\rho = \frac{\lambda}{n\mu} = \frac{20,000}{31 \cdot 480} = 1.344 > 1$$

עם תוספת של 10 רופאים נקבל: $\rho = 1.344 > 1$. תחת ההנחות הנתונות המערכת אינה יציבה, גם לא עם תוספת של 10 רופאים. בפועל, בשל העובדה שחלק מהחולים "נוטשים" (למשל, נפטרים במהלך המתנה לטיפול בבקשה) המערכת אינה מתפוצצת. אך כמובן שמצב זה אינו רצוי.

3.4 למעשה, חלק מהחולים "נוטשים" את המערכת בטרם הגעת בקשתם לטיפול (יכולות להיות לכך סיבות שונות כגון מציאת תחלופה אחרת, שיפור במצבם ולמרבה הצער גם פטירה). נמדל כעת את המערכת באמצעות מודל **Erlang A**. נבחן את מצב המערכת לאחר תוספת של 10 רופאים, כלומר עם איוש של 31 רופאים. נניח כי זמן השהייה הממוצע במערכת זו הינו כחצי שנה. באיזה תחום תפעולי פועלת המערכת? מהו כלל האיוש המתאים לתחום זה? נמקו את תשובתכם.

תשובה:

המערכת פועלת בתחום ה- ED : משך המתנה ממוצע גדול בסדרי גודל ממשך שירות מאפיין מערכות בתחום ה- ED .
 כלל האיוש המתאים הוא: $n = R - \gamma R$, $0 < \gamma < 1$, ופרקטית $0 \leq \gamma \leq 0.25$.

3.5 מהו אחוז החולים הנוטשים את המערכת? האם התשובה לשאלה זו תומכת במסקנה שהגעתם אליה בסעיף הקודם? נמקו בקצרה את תשובתכם.

תשובה:

במודל Erlang A, בתחום ה- ED : $P\{Ab\} \approx \gamma$. כלומר כ- 25% מהחולים הממתינים לאישור עוזבים את התור במהלך ההמתנה לטיפול בבקשתם. תשובתנו כאן מחזקת את התשובה בסעיף הקודם – אחוז נוטשים גבוה (דו ספרתי) מאפיין מערכות בתחום ה- ED .

3.6 השתמשו בנתוני סעיפים 3.3-3.5 כדי לחשב את הפרמטרים של המודל :

תשובה :

$$\lambda = 20,000 \text{ requests / year} \quad \theta = \frac{1}{2} \left(\frac{1}{\theta} = 2 \text{ years} \right)$$

$$\mu = 480 \text{ requests / year} \quad n = 31$$

את פרמטר הסבלנות θ חישבנו על פי הקשר : $P\{Ab\} = \theta E(W_q)$ וקיבלנו :

$$\theta = \frac{P\{Ab\}}{E(W_q)} = \frac{0.25}{0.5} = 0.5 . \text{ המשמעות : "סבלנות" ממוצעת של שנתיים.}$$

הערה : זמן טיפול רופא בבקשה נמדד בימים, וזמן שהייה במערכת בחודשים. ההנחה שזמן

$$\text{השהייה במערכת הוא } 1/2 \text{ שנה מצדיקה לכן } E(W_q) \approx \frac{1}{2} .$$

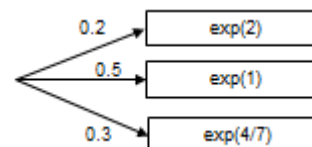
3.7 כעת נניח כי במקום 31 רופאים שיעסקו במתן אישורים לשימוש בקנאביס רפואי במשרה חלקית, בנוסף לתפקידיהם האחרים, הוחלט להקים מחלקה שתמנה 12 רופאים שיעסקו בכך במשרה מלאה.

נניח כי בכל יום עבודה מגיעות למערכת 84 בקשות (שוב, בתהליך פואסון). נניח כי ניתן לחלק את אוכלוסיית החולים המגישים בקשות לשימוש בקנאביס רפואי ל-3 סוגים עיקריים : חולים סופניים (20% מהחולים), חולי סרטן (שאינם סופניים, 50%) וחולים אחרים (30%). נניח כי, למשך הטיפול בבקשה של חולה סופני תוחלת של חצי שעה, לחולה סרטן תוחלת של שעה אחת, ולחולים האחרים תוחלת של 1.75 שעות. בסה"כ, תוחלת משך השירות הינה 1.125 שעות. משכי הטיפול מתפלגים כולם מעריכית. נניח שהקצב שבו חולים עוזבים את התור (בטרם טיפול בבקשתם) נמוך מאוד ביחס לקצב השירות, ולכן נזניח בשאלה זו את הנטישה מהתור. איזה מודל מתמטי מתאים לתיאור המערכת כעת? מהן הפרמטרים של המודל? מהן הנחות המודל? תארו את המודל והפרמטרים שלו במדויק.

תשובה :

המודל המתמטי המתאים הוא מודל M/PH/n. הנחות המודל והפרמטרים :

- הגעות לפי תהליך פואסון עם קצב של $\lambda = 84$ בקשות ליום.
- משכי שירות מפורגים PH (Hyper-exponential) עם ממוצע של 1.125 שעות :



- 12 שרתים.
- כמו כן מניחים כי אין נטישות ולא קיימת תלות בין תהליך ההגעה למשכי השירות.

שימו לב: דפי הרצאה ותרגול רלוונטיים לפתרון הסעיפים הבאים מופיעים בסוף השאלה בעמודים 22-28.

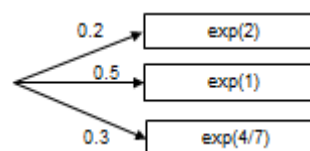
3.8 האם תוחלת זמן ההמתנה במערכת זו זהה/קטנה/גדולה מתוחלת זמן ההמתנה במערכת שבה קיימת אוכלוסייה אחת עם משך שירות המתפלג מעריכית עם ממוצע של 1.125 שעות? נמקו את תשובתכם. האם ניתן להשיב על שאלה זו גם ללא חישוב מספרי?

תזכורת: אם $X \sim \exp(\lambda)$ אזי $E[X^2] = (E[X])^2 + \text{Var}(X) = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2}$

תשובה:

כאמור, על פי סעיף 3.7 משך השירות מתפלג (Hyper-exponential) PH. ניתן לתאר את משך

השירות באופן הבא:



נחשב את שונות משך השירות:

$$E[S^2] = 0.2 \left[\{\exp(2)\}^2 \right] + 0.5 \left[\{\exp(1)\}^2 \right] + 0.3 \left[\{\exp(4/7)\}^2 \right]$$

$$= 0.2 \cdot \frac{2}{2^2} + 0.5 \cdot \frac{2}{1^2} + 0.3 \cdot \frac{2}{(4/7)^2} = 2.9375$$

$$\text{Var}(S) = E[S^2] - (ES)^2 = 2.9375 - 1.125^2 = 1.67$$

על פי Allen-Cunneen Approximation המנה בין תוחלת זמן ההמתנה בתור זה לתוחלת זמן

ההמתנה בתור M/M/n עם תוחלת משך שירות זהה היא: $\frac{1+C^2(S)}{2}$ כאשר

$$C^2(S) = \frac{\text{Var}(S)}{(E(S))^2} \quad \text{נחשב את הערך של מנה זו:}$$

$$\frac{1+C^2(S)}{2} = \frac{1+1.32}{2} = 1.16 > 1, \quad C^2(S) = \frac{1.67}{1.125^2} = 1.32 > 1$$

ולכן ממוצע זמן ההמתנה בתור זה יהיה גדול ממוצע זמן ההמתנה בתור M/M/n.

תוצאה זו הייתה צפויה שכן בהתפלגות Hyper-exponential, $C > 1$; כלומר, יש יותר אקראיות במערכת ולכן, עם אותו מספר שרתים (ועם אותה תוחלת משך שירות ואותו קצב הגעה) נקבל משכי המתנה ארוכים יותר.

3.9 החוק החדש קובע יעד של מתן מענה לחולים סופניים תוך 48 שעות. נניח כי 12 הרופאים

המוסמכים למתן אישורים מפוזרים ברחבי הארץ, כל אחד מקבל $\frac{1}{12}$ מהבקשות באקראי. כמו

כן נניח כי ביום עבודה יש 8 שעות וכי כניסת הבקשות מתפרסת באופן אחיד על פני היום. הניחו כי בקשות של חולים סופניים מקבלות עדיפות עליונה, כלומר בקשה של חולה סופני שנכנסת לתור מקבלת עדיפות על פני בקשות של חולים שאינם סופניים שנמצאות בתור, עם זאת רופא אינו מפסיק את הטיפול בבקשה בשל כניסת בקשה של חולה סופני. האם אתם **מעריכים** כי תחת האיוש הנתון ניתן יהיה לעמוד ביעד? האם ניתן יהיה לעמוד ביעד זה גם ללא ניהול עדיפויות? הביאו חישובים **התומכים** במסקנתכם, והציעו הצעות לייעול המערכת במידה ואינה עומדת ביעדיה.

שימו לב, בעמוד הבא יש מקום נוסף לענות על השאלה.

תשובה:

מדובר בתור עדיפויות Non-preemptive, כאשר בקשות של חולים סופניים מקבלות את העדיפות הגבוהה ביותר. נסתכל על הבקשות המגיעות לכל רופא כעל תור M/G/1. נחשב את תוחלת זמן ההמתנה של בקשות של חולים סופניים. קצב כניסה כללי של בקשות הינו:

$$\lambda = \frac{84}{8 \cdot 12} = 0.875 \text{ requests / hour (per doctor)}$$

$$\lambda_1 = 0.2 \cdot 0.875 = 0.175 \text{ requests / hour : קצב כניסת בקשות של חולים סופניים}$$

$$\lambda_2 = 0.5 \cdot 0.875 = 0.4375 \text{ requests / hour : קצב כניסת בקשות של חולי סרטן}$$

$$\lambda_3 = 0.3 \cdot 0.875 = 0.2625 \text{ requests / hour : קצב כניסת בקשות של חולים אחרים}$$

$$\mu_1 = \frac{1}{0.5} = 2 \text{ : קצב טיפול בבקשות של חולים סופניים}$$

$$\mu_2 = 1 \text{ : קצב טיפול בבקשות של חולי סרטן}, \mu_3 = 4/7 \text{ : קצב טיפול בבקשות של חולים אחרים}$$

מכיוון שמשכי השירות של כל סוגי הבקשות מתפלגים מעריכית, $C_k^2(S) = 1$ עבור כל סוגי הבקשות. מכאן, תוחלת שארית משך השירות של בקשה הנמצאת בטיפול בעת הגעת בקשה של חולה סופני הינה:

$$E(R) = \rho_1 E(S_1) + \rho_2 E(S_2) + \rho_3 E(S_3) = \frac{0.175}{2} \cdot 0.5 + \frac{0.4375}{1} \cdot 1 + \frac{0.2625}{(4/7)} \cdot 1.75 = 1.285$$

$$E(W_q^1) = \frac{E(R)}{1 - \rho_1} = \frac{1.285}{1 - 0.0875} = 1.408 \text{ hours ולכן}$$

יחד עם זמן הטיפול בבקשה נקבל כי בקשה של חולה סופני מטופלת בממוצע תוך קצת יותר משעתיים מכניסתה למערכת.

זוהי אמנם רק תוחלת זמן המתנה, ולכן יש גם חולים עבורם זמן ההמתנה יהיה גבוה יותר. אך ניתן להניח כי אם התוחלת כל-כך נמוכה, מרבית הבקשות של החולים הנ"ל יקבלו מענה תוך 48 שעות.

אם לא ננהל תור עדיפויות תוחלת זמן ההמתנה של כל החולים תהיה :

$$E(W_q) \approx E(S) \frac{\rho}{1-\rho} \cdot \frac{1+C^2(S)}{2} = 1.125 \cdot \frac{0.984}{1-0.984} \cdot \frac{1+1.32}{2} = 80.25 \text{ hours}$$

מהיעד שנקבע.

אם נשנה את ארכיטקטורת המערכת מ-12 מערכות מ-12 מערכות של $M / G / 1$, למערכת אחת של $M / G / 12$ (מערכת מידע תנהל תור יחיד עבור כל 12 הרופאים), ממוצע זמן המתנה בהינתן המתנה יקוצר פי 12, ואז נעמוד ביעדים.

3.10 מטרה נוספת של משרד הבריאות היא למזער את מספר החולים שנפטרים בטרם בקשתם מגיעה לטיפול. נניח כי בין שלושת סוגי החולים ישנם הבדלים גם בתוחלת משך החיים לאחר פתיחת הבקשה. נסמן ב- τ_i את תוחלת החיים (לאחר פתיחת בקשה) של חולים מסוג i . בהשתמשכם במה שלמדתם בקורס, הציעו כלל אינטואיטיבי לקביעת עדיפויות בסדר קבלת החולים שעשוי להביא להשגת מטרה זו. הסבירו את תשובתכם.

תשובה :

נסמן ב- μ_i את משך הטיפול בבקשות של חולים מסוג i . הכלל שנציע יהיה מתן עדיפויות על פי

$$\text{ערך } \frac{\mu_i}{\tau_i} = \theta_i \mu_i, \text{ כלומר ניתן עדיפות גבוהה ביותר לחולים עבורם ערך זה הוא הגבוה ביותר}$$

ועדיפות נמוכה ביותר לחולים עבורם ערך זה הוא הנמוך ביותר.

באופן אינטואיטיבי, בהקבלה לכלל ה- $c\mu$ שנלמד בכיתה, אם נניח כי קצבי הטיפול זהים,

כלומר $\mu_i = \mu \quad \forall i$ אזי ניתן עדיפות לבקשות של חולים בעלי תוחלת החיים הקצרה ביותר, כלומר

θ_i הגבוה ביותר.

אם נניח שתוחלת החיים שווה אזי ניתן שירות לפי כלל ה- SPT (Shortest Process Time)

ונקבל קודם את הבקשות של החולים שעבורם משך הטיפול הוא הקצר ביותר, כלומר μ_i הוא

הגבוה ביותר.

ואם גם קצבי השירות שונים וגם תוחלת החיים שונה ניתן עדיפות על פי הכלל שצוין לעיל.

M/M/n Queue: Properties

Erlang-C Formula (1917) for the Delay probability:

$$P\{W_q > 0\} \triangleq E_{2,n} = \sum_{i \geq n} \pi_i = \frac{R^n}{n!} \frac{1}{1-\rho} \cdot \pi_0.$$

Erlang-C computation: via recursion, see Erlang-B below.

Erlang-C approximations: important later in course.

Number-in-queue:

$$P\{L_q = i\} = E_{2,n} \cdot (1-\rho)\rho^i, \quad i > 0,$$

or

$$L_q = \begin{cases} 0 & \text{wp } 1 - E_{2,n} \\ \text{Geom}_{\geq 0}(1-\rho) & \text{wp } E_{2,n} \end{cases}$$

Waiting time distribution:

$$\frac{W_q}{1/\mu} = \begin{cases} 0 & \text{wp } 1 - E_{2,n} \\ \text{Exp}(\text{mean} = \frac{1}{n} \cdot \frac{1}{1-\rho}) & \text{wp } E_{2,n} \end{cases}$$

Compare with M/M/1!

Departure process: Poisson(λ) in steady-state.

Proof via reversibility, as with M/M/1.

M/M/1: Further Properties

- **Delay probability** (PASTA):

$$P\{W_q > 0\} = \rho.$$

- **Waiting time in queue** (given delay, it is exp):

$$\frac{W_q}{1/\mu} \stackrel{d}{=} \begin{cases} 0 & \text{wp } 1 - \rho \\ \exp\left(\text{mean} = \frac{1}{1-\rho}\right) & \text{wp } \rho \end{cases}$$

Note: $E[\frac{W_q}{1/\mu}] = 0 \times (1 - \rho) + \frac{1}{1-\rho} \times \rho = \frac{\rho}{1-\rho}.$

- **Number-in-system/queue:**

$$E[L] = \frac{\rho}{1-\rho}; \quad E[L_q] = \frac{\rho^2}{1-\rho}.$$

- **Server's utilization** (occupancy) is $\rho = \lambda/\mu$.

Via Little's formula, applied to "system = server":

$$\rho = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T L_{Server}(u) du = \lambda \times \frac{1}{\mu}.$$

- **Departure process** in steady state is Poisson (λ) (Burke's theorem) – useful in queueing networks.

Support: Reversibility implies that the departure process equals (in distribution) the arrival process. Furthermore,

Average inter-departure time =

$$\frac{1}{\mu} \cdot \rho + \left(\frac{1}{\lambda} + \frac{1}{\mu}\right) \cdot (1 - \rho) = \frac{1}{\lambda}.$$

Approximating G/G/n

Stability condition: $\rho = \frac{\lambda}{n\mu} < 1$.

Kingman's Exponential Law:

$$\frac{W_q}{E(S)} \approx \begin{cases} \exp\left(\text{mean} = \frac{1}{n} \cdot \frac{1}{1-\rho} \cdot \frac{C^2(A)+C^2(S)}{2}\right) & , \text{ wp } E_{2,n}, \\ 0 & , \text{ otherwise.} \end{cases}$$

In particular, a popular measure for service-level, used to determine the number-of-servers n , is:

$$P\{W_q > x \cdot E(S)\} \approx E_{2,n} \cdot \exp\left(-x \cdot \frac{2n(1-\rho)}{C^2(A) + C^2(S)}\right), \quad x > 0.$$

Allen-Cunneen Approximation:

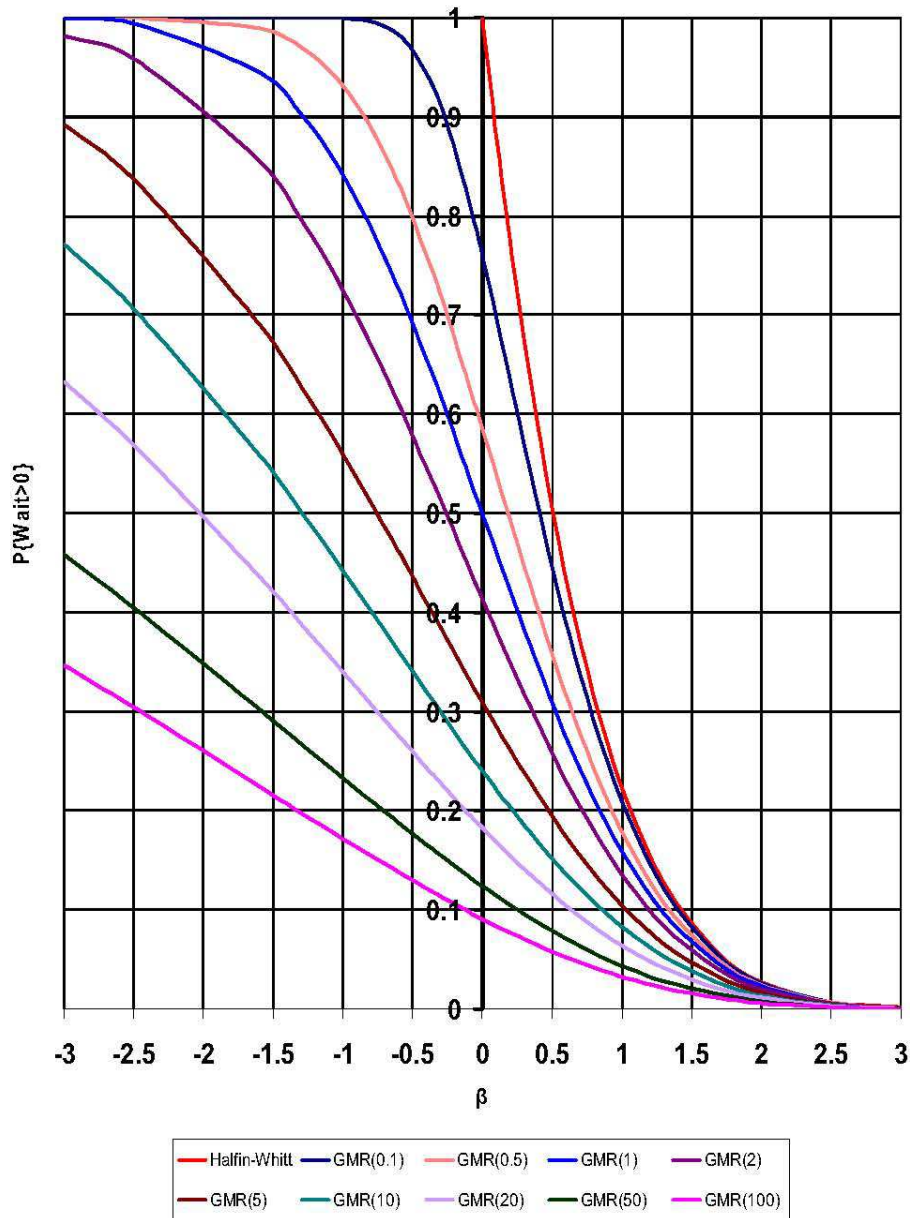
$$E(W_q) \approx E(S) \cdot \frac{1}{n} \cdot \frac{E_{2,n}}{1-\rho} \cdot \frac{C^2(A) + C^2(S)}{2}.$$

or equivalently,

$$E(W_q) \approx E(W_{q,M/M/n}) \cdot \frac{C^2(A) + C^2(S)}{2}.$$

Erlang-A: The Garnett Delay-Functions

$P\{W_q > 0\}$ vs. the QOS parameter β , for varying patience θ/μ .



GMR(x) describes the asymptotic probability of delay as a function of β when $\theta/\mu = x$. Here, θ and μ are the abandonment and service rate, respectively.

Note: **Erlang-C** = limit of **Erlang-A**, as patience $\uparrow$ indefinitely.

Recitation 13: Priority Queues

M/G/1 with priorities

- K customer classes, indexed by $k = 1, \dots, K$.
- **Class k arrivals:** Poisson, rate λ_k .
- **Class k service times:** S_k - generally distributed, with $m_k = E(S_k)$ and $E(S_k^2)$ both finite.
- **Setting the priorities:** Set highest priorities to 1, then $2, \dots$; lowest to K .
- Assume FCFS within each priority class.
- Non preemptive first (Later, preemptive-resume).

Steady state $\Leftrightarrow \rho \triangleq \rho_1 + \dots + \rho_K < 1$, where $\rho_k = \lambda_k m_k$.

Convenient notation: $\bar{\rho}_k = \rho_1 + \dots + \rho_k$, $1 \leq k \leq K$.

Note: ρ_k = fraction of time allocated by server to class k .
 $1 - \rho$ = idleness/availability.

- * $E(W_q^k)$ - expected waiting time of class k customer.
- * $E(L_q^k)$ - expected number of waiting class k customers.
- * $E(U)$ - expected unfinished work in the system.
- * $E(R)$ - expected residual service time.

Calculation of $E(W_q^k)$. Non-preemptive regime

$$\begin{aligned}
 1. \quad & E(W_q^1) = E(R) + m_1 E(L_q^1) = E(R) + \rho_1 E(W_q^1) \\
 & \Rightarrow E(W_q^1) = E(R)/(1 - \rho_1), \text{ as before } (K = 1). \\
 2. \quad & E(W_q^2) = E(R) + \underbrace{m_1 E(L_q^1) + m_2 E(L_q^2)}_{\text{wait due to class 1 \& 2 in queue}} + \underbrace{m_1 \lambda_1 E(W_q^2)}_{\text{wait due to class 1, arriving during wait of 2.}} \\
 & \Rightarrow E(W_q^2) = E(R) + \rho_1 E(W_q^1) + \rho_2 E(W_q^2) + \rho_1 E(W_q^2) \\
 & \Rightarrow E(W_q^2) = [E(R) + \rho_1 E(W_q^1)]/(1 - \rho_1 - \rho_2) = \\
 & \quad = E(R)/[(1 - \rho_1)(1 - \rho_1 - \rho_2)] \\
 & \quad \quad \quad \uparrow \\
 & \quad \quad \text{substitute } E(W_q^1)
 \end{aligned}$$

$$\begin{aligned}
 k. \quad & EW_q^k = ER + m_1 \cdot EL_q^1 + \dots + m_k \cdot EL_q^k + \lambda_1 m_1 EW_q^k + \\
 & \dots + \lambda_{k-1} m_{k-1} EW_q^k \\
 & \Rightarrow = ER + \rho_1 EW_q^1 + \dots + \rho_{k-1} EW_q^{k-1} + (\rho_1 + \dots + \rho_k) EW_q^k
 \end{aligned}$$

$$\begin{aligned}
 E(W_q^k) &= \frac{E(R) + \rho_1 E(W_q^1) + \dots + \rho_{k-1} E(W_q^{k-1})}{(1 - \rho_1 - \rho_2 - \dots - \rho_k)}, \quad k \geq 1 \\
 &= (\text{Induction}) \frac{E(R) \cdot \left[1 + \frac{\rho_1}{1 - \rho_1} + \frac{\rho_2}{(1 - \rho_1)(1 - \rho_1 - \rho_2)} + \frac{\rho_{k-1}}{(1 - \rho_{k-2})(1 - \rho_{k-1})} \right]}{1 - \bar{\rho}_k} \\
 &= \frac{E(R)}{(1 - \bar{\rho}_{k-1})(1 - \bar{\rho}_k)}
 \end{aligned}$$

The last equality can be derived via simple calculations.

We now show $E(R) = \frac{1}{2} \sum_{k=1}^K \lambda_k E(S_k^2)$

$$E(R) = (1 - \rho) \cdot 0 + \sum_k \rho_k \cdot m_k \cdot \frac{1 + C_k^2(S)}{2} = \frac{1}{2} \sum_k \lambda_k E(S_k^2)$$

$$\Rightarrow \boxed{E(W_q^k) = \frac{\frac{1}{2} \sum_{j=1}^K \lambda_j E(S_j^2)}{(1 - \rho_1 - \dots - \rho_{k-1})(1 - \rho_1 - \dots - \rho_k)}}, 1 \leq k \leq K.$$

Calculation of $E(W_q^k)$. Preemptive regime

Now, Class k *does not* “see” classes $k+1, \dots, K$.

Recall: for M/G/1-like queues, $E(U) = \frac{E(R)}{1 - \rho} = E(W_q)$

$$E(W_q^k) = \frac{E(R^k)}{1 - (\rho_1 + \dots + \rho_k)} + \sum_{j=1}^{k-1} \lambda_j m_j [E(W_q^k) + m_k]$$

$\uparrow$
 $j \leq k-1$ preempts k

$$= \frac{E(R^k)}{1 - \bar{\rho}_k} + \bar{\rho}_{k-1} [E(W_q^k) + m_k]$$

$$E(W_q^k) = \frac{E(R^k)}{(1 - \bar{\rho}_k)(1 - \bar{\rho}_{k-1})} + \frac{\bar{\rho}_{k-1}}{1 - \bar{\rho}_{k-1}} m_k$$

where $E(R^k) = \sum_{j=1}^k \rho_j \cdot m_j \cdot \frac{1 + C^2(S_j)}{2} = \frac{1}{2} \sum_{j=1}^k \lambda_j E(S_j^2)$

$$\boxed{E(W_q^k) = \frac{\frac{1}{2} \sum_{j=1}^k \lambda_j E(S_j^2)}{(1 - \bar{\rho}_{k-1})(1 - \bar{\rho}_k)} + \frac{\bar{\rho}_{k-1}}{1 - \bar{\rho}_{k-1}} E(S_k)}$$

שאלה 4 – תאוריה (16 נקודות)

בשני העמודים הבאים מופיעים דפי הרצאה : משפט ה – QED למודל Erlang B והוכחת המשפט ממנה נמחקו חלקים מסויימים.

4.1 במשפט ה – QED, מהי האינטרפרטציה המתמטית המדויקת של "הבט המנהל", דהיינו

$$n \approx R_n + \beta \sqrt{R_n}, \quad R_n = \frac{\lambda_n}{\mu}$$

תשובה :

$$\lim_{n \rightarrow \infty} \frac{n - R_n}{\sqrt{R_n}} = \beta$$

4.2 בהוכחה, השלימו את חלק A החסר.

תשובה :

$$P \left\{ \frac{X_R - R}{\sqrt{R}} \leq \frac{n - R}{\sqrt{R}} \right\}$$

4.3 בהוכחה, השלימו את חלק B החסר.

תשובה :

$$\begin{aligned} & P\{n - 1 < X_R \leq n\} \\ &= P \left\{ \frac{n - R - 1}{\sqrt{R}} < \frac{X_R - R}{\sqrt{R}} \leq \frac{n - R}{\sqrt{R}} \right\} \\ &\approx P \left\{ \beta - \frac{1}{\sqrt{R}} \leq N(0, 1) \leq \beta \right\} \end{aligned}$$

4.4 במשפט ה – QED, הוכיחו את C המופיע במסגרת, דהיינו הוכיחו כי $\frac{\phi(\beta)}{\Phi(\beta)} = h(-\beta)$,

כאשר ϕ זוהי פונקציית הצפיפות של משתנה מקרי $N(0,1)$ ו- Φ פונקציית ההתפלגות המצטברת של אותה התפלגות.

תשובה :

מסימטריה של התפלגות נורמלית סטנדרטית $\phi(\beta) = \phi(-\beta)$ וכן $\Phi(\beta) = 1 - \Phi(-\beta)$.

$$\text{ולכן : } \frac{\phi(\beta)}{\Phi(\beta)} = \frac{\phi(-\beta)}{1 - \Phi(-\beta)} = h(\beta)$$

The Erlang-B Queue in the QED-Regime

Recall the **Erlang-B Formula**:

$$E_{1,n} \triangleq P\{\text{Blocked}\} = \frac{R^n}{n!} \bigg/ \sum_{j=0}^n \frac{R^j}{j!}$$

Consider a sequence of **M/M/n/n** queues,
indexed by **the number of servers $n = 1, 2, \dots$** .

- λ_n = arrival-rate, varies with n ;
- μ = service-rate, fixed (independent of n).
- $R_n = \lambda_n/\mu$ (Offered Load) ; $\rho_n = R_n/n$ (Load per Server);
We shall use R and ρ , without the subscript n , for simplicity.

Theorem (QED Erlang-B; Jagerman, 1974)

As $n \rightarrow \infty$, the following 3 statements are equivalent:

1. **Customers:** $E_{1,n} \approx \frac{\gamma}{\sqrt{n}}$, for some $\gamma > 0$;
2. **Servers:** $\rho \approx 1 - \frac{\beta}{\sqrt{n}}$, for some $-\infty < \beta < \infty$;
3. **Manager:** $n \approx R + \beta\sqrt{R}$ (square-root “staffing”);

in which case

$$\gamma = \boxed{h(-\beta) = \frac{\phi(-\beta)}{\bar{\Phi}(-\beta)} = \frac{\phi(\beta)}{\Phi(\beta)}, \quad \mathbf{C}} \longleftarrow \text{הוכיחו:}$$

where $\phi, \Phi, \bar{\Phi}$ and h are the density, cdf, survival function and hazard rate of $N(0, 1)$ (standard-normal), respectively.

Note: **Servers' Occupancy** $\approx 1 - \frac{\beta+\gamma}{\sqrt{n}}$, accounting for blocking.

QED Erlang-B: Proof

Proof:

2 $\iff$ **3** is straightforward algebra from the definitions.

3 $\Rightarrow$ **1**. Assume $n = R + \beta\sqrt{R}$. The key observation is a **Poisson-Representation** of the Erlang-B Formula:

$$E_{1,n} = \frac{P\{X_R = n\}}{P\{X_R \leq n\}},$$

where $X_R \stackrel{d}{=} \text{Poisson}(R)$.

$$P\{X_R \leq n\} = \boxed{\mathbf{A}} \xleftarrow{\text{השלימו:}} \underset{CLT,3}{\approx} P\{N(0,1) \leq \beta\} = \Phi(\beta).$$

$$P\{X_R = n\} = \boxed{\mathbf{B}} \xleftarrow{\text{השלימו:}} \\ \approx \frac{1}{\sqrt{R}} \cdot \phi(\beta) \approx \frac{1}{\sqrt{n}} \cdot \phi(\beta).$$

Table 1: Discrete and Continuous Distributions

Distribution	Notation	Probability Function	Parameters	Support	Moment Gen. Function	Expectation	Variance	Comments
Binomial	$N \sim \text{Bin}(n, p)$	$p(k) = \binom{n}{k} p^k q^{n-k}$	$0 \leq p = 1 - q \leq 1; n \geq 1$	$k = 0, 1, \dots, n$	$[p \cdot e^t + q]^n$	np	npq	
Multi-nomial	$X \sim \text{Mult}(n, p)$	$p(x) = \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^k p_i^{x_i}$ $x = (x_1, \dots, x_k)$	$0 \leq p_i; \sum_{i=1}^k p_i = 1; n \geq 1$ $p = (p_1, \dots, p_k)$	$0 \leq x_i; \sum_{i=1}^k x_i = n$		$E(X_i) = np_i$ $i = 1, \dots, k$	$\text{cov}(X_i, X_j) = -np_i p_j$ $(i \neq j)$	$X_i \sim \text{Bin}(n, p_i)$
Negative Binomial	$X \sim NB(m, p)$	$p(k) = \binom{k-1}{m-1} p^m q^{k-m}$	$0 \leq p = 1 - q \leq 1; m \geq 1$	$k = m, m+1, \dots$	$\left[\frac{p \cdot e^t}{1 - q \cdot e^t} \right]^m$	$\frac{m}{p}$	$\frac{mq}{p^2}$	Number of trials till m -th success
Geometric	$X \sim G(p)$	$p(k) = q^{k-1} \cdot p$	$0 \leq p \leq 1$	$k = 1, 2, \dots$	$\frac{p \cdot e^t}{1 - q \cdot e^t}$	$\frac{1}{p}$	$\frac{q}{p^2}$	$X \sim NB(1, p)$
Poisson	$X \sim \mathcal{P}(\lambda)$	$p(k) = \frac{e^{-\lambda} \lambda^k}{k!}$	$\lambda \geq 0$	$k = 0, 1, \dots$	$\exp\{-\lambda(1 - e^t)\}$	λ	λ	
Hyper-geometric	$X \sim HG(N, M, n)$	$p(k) = \frac{\binom{N}{k} \binom{M}{n-k}}{\binom{N+M}{n}}$	$N, M > 0;$ $1 \leq n \leq N + M$	$\max(0, n - M) \leq k$ $k \leq \min(n, N)$	$\sum_k e^{tk} p(k)$	$\frac{n \cdot N}{N+M}$	$n \cdot \frac{N}{N+M} \cdot \frac{M}{N+M} \cdot \frac{N-M-n}{N+M-1}$	N – special elements M – ordinary ones n – sample size k – special in sample
		Density Function						
Beta	$X \sim \text{Beta}(\alpha, \beta)$	$f(t) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} t^{\alpha-1} (1-t)^{\beta-1}$	$\alpha, \beta > 0$	$0 \leq t \leq 1$	Complicated	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\alpha = \beta = 1 \rightarrow \text{Uniform}$
Gamma	$X \sim \text{Gamma}(\alpha, \lambda)$	$f(t) = \frac{\lambda^\alpha t^{\alpha-1} e^{-\lambda t}}{\Gamma(\alpha)}$	$\alpha, \lambda > 0$	$t \geq 0$	$(1 - \frac{t}{\lambda})^{-\alpha}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$ $= (x-1)\Gamma(x-1)$
Erlang	$X \sim \text{Erlang}(n, \lambda)$	$f(t) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}$	$n = 1, 2, \dots; \lambda > 0$	$t \geq 0$	$(1 - \frac{t}{\lambda})^{-n}$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$	$X \sim \text{Gamma}(n, \lambda)$
Exponential	$X \sim \text{Exp}(\lambda)$	$f(t) = \lambda e^{-\lambda t}$	$\lambda > 0$	$t \geq 0$	$(1 - \frac{t}{\lambda})^{-1}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$X \sim \text{Gamma}(1, \lambda)$
Normal	$X \sim N(\mu, \sigma^2)$	$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}$	$-\infty < \mu < \infty; \sigma > 0$	$-\infty < t < \infty$	$\exp\left\{t\mu + \frac{1}{2}t^2\sigma^2\right\}$	μ	σ^2	