



מרצה : פרופסור אבישי מנדלבאום

מתרגלת : ניצן כרמלי

תאריך הבחינה : 21.03.2012

מס' סטודנט _____

שם _____

הנדסת מערכות שירות – 096324

מועד ב' – סמסטר חורף תשע"ב 2012

פתרון

- המבחן נמשך שלוש שעות. באחריותך להחזירו במועד.
- המבחן בחומר סגור.
- מותר להשתמש במחשבון.
- המבחן כולל 30 עמודים, כולל עמוד זה.
- 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 10 נקודות	שאלה 1	ניקוד:
תרגילי בית	מתוך 10 נקודות	שאלה 2	
יישום	מתוך 20 נקודות	שאלה 3	
תיאוריה	מתוך 10 נקודות	שאלה 4	

סה"כ _____ מתוך 50 נקודות

בהצלחה !

שאלה 1 - תרגילי כיתה/הרצאות/מבחנים קודמים (10 נקודות)

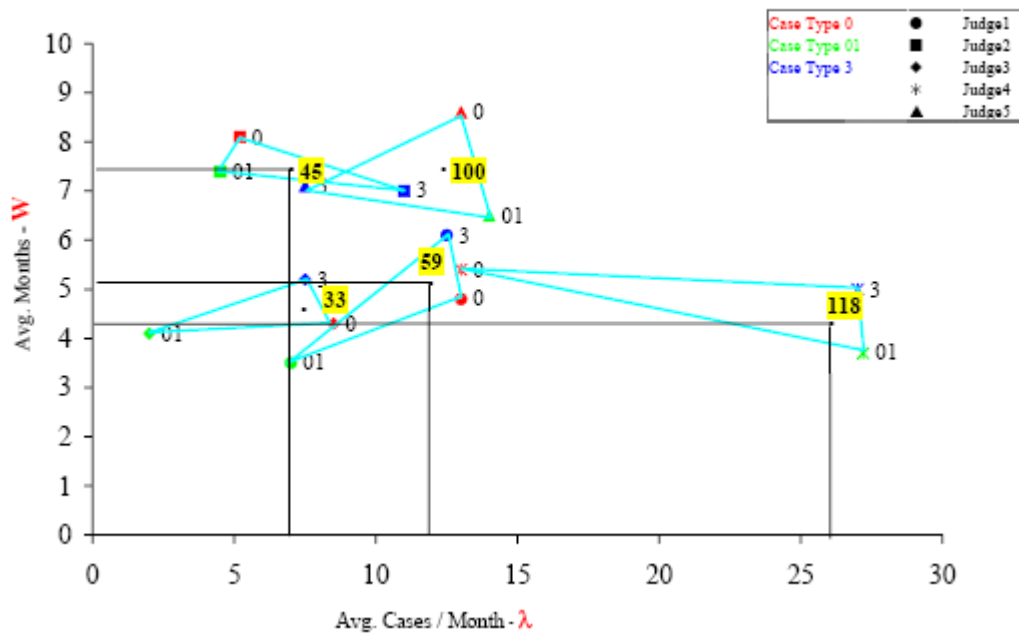
חלק 1

בכל אחת מהשאלות הבאות, בחרו את התשובה הנכונה (אחת מארבע האפשרויות).

1.1.1

גרף 1

Judges: The Best/Worst (Operational) Performer



התרשים הנ"ל שהוצג בהרצאה מתאר את רמת הביצועים התפעוליים של כמה שופטים בבית הדין לעבודה. שטחי 3 המלבנים שאנו רואים בציור מתארים:

1. מספר תיקים ממוצע ששופט מסיים במהלך חודש.

2. מספר תיקים ממוצע שנמצאים בטיפול אצל שופט בעת ובעונה אחת.

3. זמן שהיה ממוצע של תיק במערכת.

4. מספר תיקים ממוצע ששופט פותח במהלך חודש.

תשובה: 2.

1.1.2 ניתן להשוות בין רמת הביצועים התפעוליים של שלושה שופטים על סמך שטחי שלושת המלבנים שאנו רואים בתרשים (השטחים שווים ל-45, 59 ו-118). מהי מהות הקשר בין שטח המלבן ורמת הביצועים של השופט?

1. שופט עם שטח מלבן גדול יותר הוא שופט עם רמת ביצועים גבוהה יותר.

2. שופט עם שטח מלבן גדול יותר הוא שופט עם רמת ביצועים נמוכה יותר.

3. לא נכון להשוות בין רמות הביצועים התפעוליים של שופטים על סמך שטחי המלבנים. לשופט עם רמת ביצועים גבוהה צריך להיות גבוה ו- W נמוך.

4. אף תשובה לא נכונה.

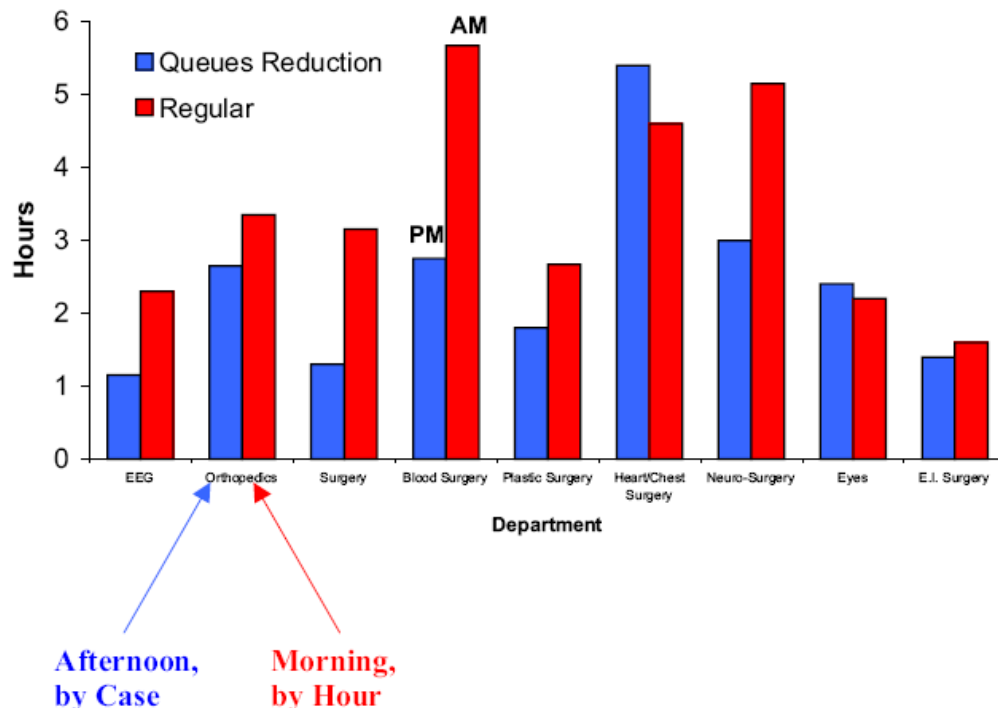
תשובה: 3.

מקום זה
הושאר ריק

1.1.3 גרף 2 שהוצג בהרצאה מתאר משכי ניתוחים בבית חולים מסוים בבוקר ואחרי הצהריים. אנו רואים שלכמה סוגי ניתוחים, זמני ניתוחים בבוקר הם הרבה יותר ארוכים מאחה"צ. בחרו פיסקה שמסכמת בצורה הסבירה ביותר את הסיבות לתופעה ואת השלכותיה הניהוליות.

גרף 2

Operations Time - Morning vs. Afternoon:

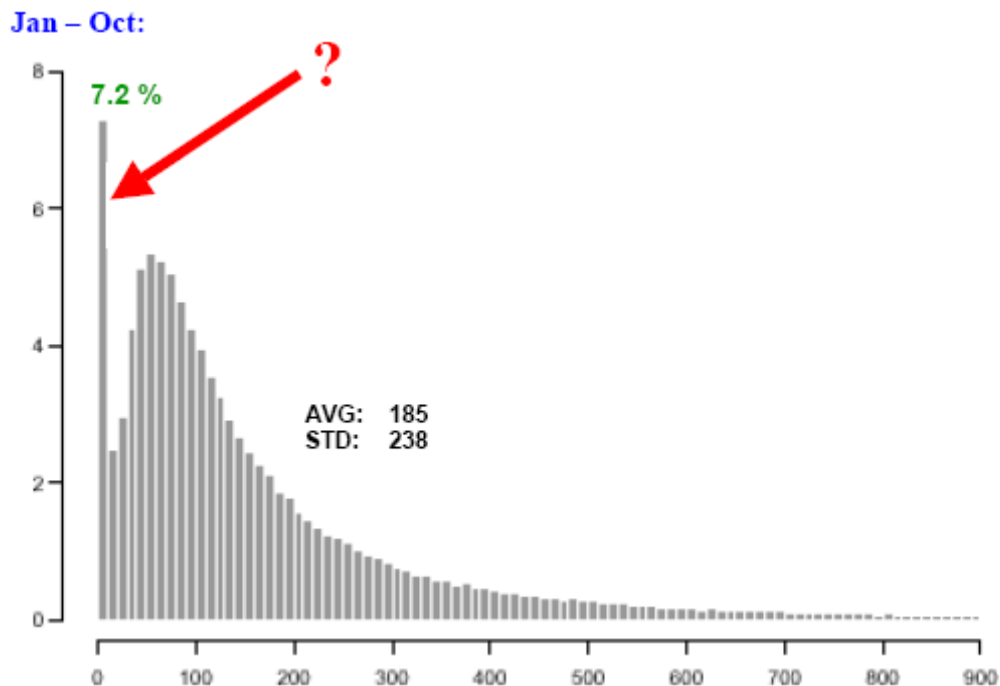


1. התופעה היא בעייתית. בגלל מערכת תמריצים לא נכונה חולים מקבלים שירות ברמה נמוכה אחה"צ, הבאה לידי ביטוי בניתוחים קצרים.
2. התופעה היא חיובית במקרה זה. מכיוון שאחה"צ המנתחים מקבלים תשלום פר ניתוח, הם משאירים ניתוחים קלים יותר למשמרת אחרי הצהריים. כתוצאה מכך, המנתחים מקדישים לניתוחים הקשים יותר את שעות הבוקר, בהם הם ערניים יותר והמערכת ערוכה יותר מאשר בשעות אחה"צ, הערב או הלילה.
3. התופעה היא בעייתית. מכיוון שבבוקר המנתחים מקבלים תשלום פר שעה, הם מבצעים פחות מדי ניתוחים בבוקר שמביא לעומס יתר בשעות אחה"צ.
4. התופעה היא חיובית במקרה זה. מערכת תמריצים מתוחכמת מאפשרת למזער הוצאות משכורת הרופאים וליעל את תהליכי ניהול בבית חולים.

תשובה : 2

1.1.4 גרף 3 לעיל מתאר היסטוגרמה של משכי השירות במוקד שירות טלפוני בארץ. אנו רואים אחוז גדול יחסית של שיחות קצרות מאוד במוקד זה. מה הן סיבות פוטנציאליות של תופעה זו?

גרף 3



1. במוקד זה 7.2% מהלקוחות נוטשים את התור הטלפוני עקב חוסר סבלנות.
2. תמריץ המבוסס על מספר השיחות המטופלות גורם למוקדנים לנתק בתחילת השיחה כדי לשפר את "רמת ביצועיהם".
3. מוקדנים נמצאים תחת לחץ עבודה כבד ומנתקים מיד עם תחילת השיחה (ובטרם נוצר קשר דיבור בין המוקדן ללקוח) כדי לנוח.
4. גם 2 וגם 3 יכולים להיות סיבות פוטנציאליות לתופעה המתוארת.

תשובה: 4.

1.1.5 את חווית ההמתנה של לקוח במרכז טלפוני סיכמנו באופן הבא :

1. הזמן שלקוח מצפה להמתין - E
2. הזמן שלקוח מוכן להמתין - τ
3. הזמן שלקוח נדרש להמתין - V
4. הזמן שלקוח למעשה המתין - W
5. הזמן שלקוח הרגיש שהמתין - F

1.1.5.1 מודל $Erlang-A$ דורש כקלט את הממוצע של רק אחד מחמשת הזמנים הנ"ל : מהו?

תשובה : 2.

1.1.5.2 מודל $Erlang-A$ נותן כפלט התפלגות של שניים מחמשת הזמנים הנ"ל : מהם?

תשובה : 3 ו- 4.

חלק 2 - חוק Little

בחברה יפנית כלשהי 36,000 עובדים, 20% מהם נשים. משך העסקה ממוצע של נשים הוא 37 חודשים, ולגברים 200 חודשים. הנח כי סך כמות המועסקים ואחוזי הגברים והנשים בתוכם קבועים לאורך זמן.

1.2.1 הסבירו מדוע קצב ההצטרפות של נשים חדשות לחברה הוא גדול יותר מקצב הצטרפות הגברים, למרות שפרופורציות הגברים והנשים במצב יציב הן 80% גברים ו-20% נשים. גבו את תשובתכם בחישובים.

תשובה :

צריך לחשב את λ_i (לגברים $i=1$, לנשים $i=2$). ידוע על פי הנתונים ש-

$$L_1 = 36,000 * 0.8 = 28,800, L_2 = 36,000 * 0.2 = 7,200$$

ולכן ע"ס חוק Little מתקיים $\lambda_1 = \frac{L_1}{W_1} = \frac{28,800}{200} = 144$, $\lambda_2 = \frac{L_2}{W_2} = \frac{7,200}{37} = 194.6$. כלומר,

מכיוון שזמני השהיה של גברים הרבה יותר גדולה משל נשים, קצב ההגעה של נשים צריך להיות גבוה יותר (אם רוצים לשמור על 20% נשים בחברה).

1.2.2 נסמן את ממוצעי זמני ההעסקה של גברים ונשים על ידי W_1 ו- W_2 בהתאמה.

נסמן ב- W את ממוצע זמן ההעסקה הכללית בחברה.

בסעיף זה נניח כי נתונים (ע"י מדידות או מנסיון) p_1 ו- p_2 : פרופורציות מספר העובדים

שנשכרים בחודש מסוג 1 (גברים) וסוג 2 (נשים). ז"א, $p_i = \frac{\lambda_i}{\lambda_1 + \lambda_2}$, כאשר λ_i הוא קצב הגעת

עובדים חדשים לחברה מסוג i , $(i = 1, 2)$.

בטאו את W בעזרת W_1 , p_2 , p_1 ו- W_2

$$W = p_1 W_1 + p_2 W_2 \quad \text{תשובה:}$$

דרך חישוב: בעזרת נוסחת Little.

$$L = L_1 + L_2 \Rightarrow \lambda W = \lambda_1 W_1 + \lambda_2 W_2 \Rightarrow W = p_1 W_1 + p_2 W_2$$

ז"א W הוא ממוצע חשבונני של W_1 ו- W_2 .

1.2.2 בסעיף זה נניח כי נתונים q_1 ו- q_2 : פרופורציות מספר העובדים המועסקים בחברה,

מסוג 1 (גברים) וסוג 2 (נשים).

ז"א, $q_i = \frac{L_i}{L_1 + L_2}$, כאשר L_i הוא ממוצע מספר העובדים בחברה מסוג i , $(i = 1, 2)$.

בטאו את W בעזרת q_2 , q_1 , W_1 ו- W_2 .

$$W = \frac{1}{\frac{q_1}{W_1} + \frac{q_2}{W_2}} \quad \text{תשובה:}$$

דרך חישוב:

$$W = \frac{L_1 + L_2}{\lambda_1 + \lambda_2} = \frac{L_1 + L_2}{\frac{L_1}{W_1} + \frac{L_2}{W_2}} = \frac{L_1 + L_2}{\frac{q_1(L_1 + L_2)}{W_1} + \frac{q_2(L_1 + L_2)}{W_2}} = \frac{1}{\frac{q_1}{W_1} + \frac{q_2}{W_2}}$$

ז"א W הוא ממוצע הרמוני של W_1 ו- W_2 .

שאלה 2 – תרגילי בית (10 נקודות)

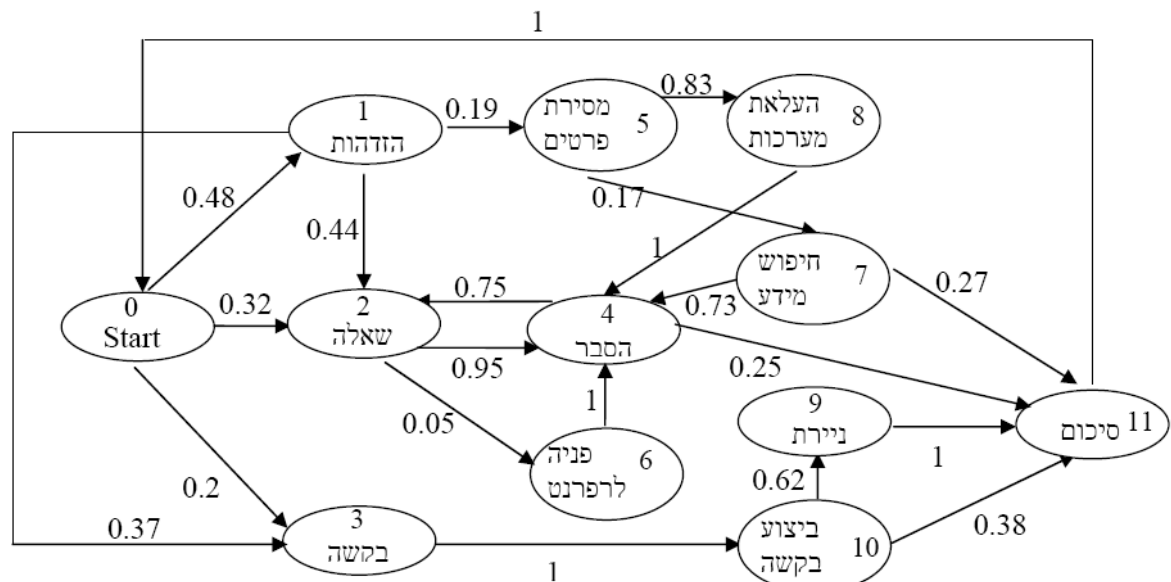
חלק 1 - התפלגות זמני שהיה

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:

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

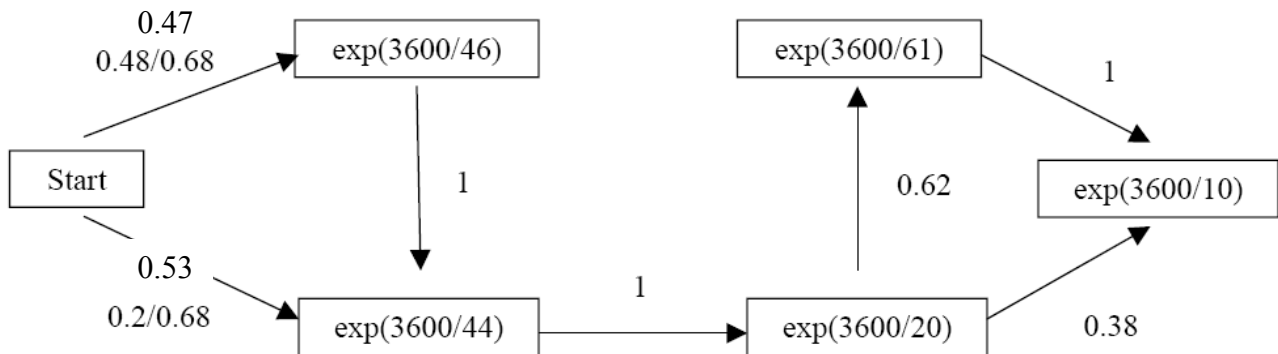
2.1.2. כיצד מתפלג משך שיחה? תארו במדויק ככל שתוכלו.

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

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

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":

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

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 פונקצית הקצב $\lambda(t)$. המוקד פועל בשעות 16:00-24:00. נסמן אינטרוול זמן זה ב- $[0, T]$.

ממוצע זמן 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"ל לדקה ועלות המתנה הינה h ש"ל לדקה. עוד הניחו כי הרווח מסיום שירות הינו r ש"ל והקנס על נטישת לקוח הינו s ש"ל.

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:

$$\frac{dQ(t)}{dt} = \lambda(t) - \mu(Q(t) \wedge N(t)) - \theta(Q(t) - N(t))^+$$

2.2.2 תארו בקיצור איך ניתן, בעזרת גיליון אלקטרוני בלבד, לפתור בקירוב (מדויק ככל הנדרש) את המשוואה הדיפרנציאלית בסעיף 1.1.1.

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 את הביטוי לרווח הכולל של המערכת $C^{(N)}$.

תשובה:

$$C^{(N)} = \int_0^T [r\mu(Q_t \wedge N_t) - s\theta(Q_t - N_t)^+ - cN_t] dt.$$

2.2.4 הניחו כעת כי מספר השרתים N קבוע במהלך יום עבודה, כאשר N^* הוא מספר השרתים האופטימאלי שיביא למקסימום רווח של המוקד. 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:

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

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

שאלה 3 – יישום (20 נקודות)

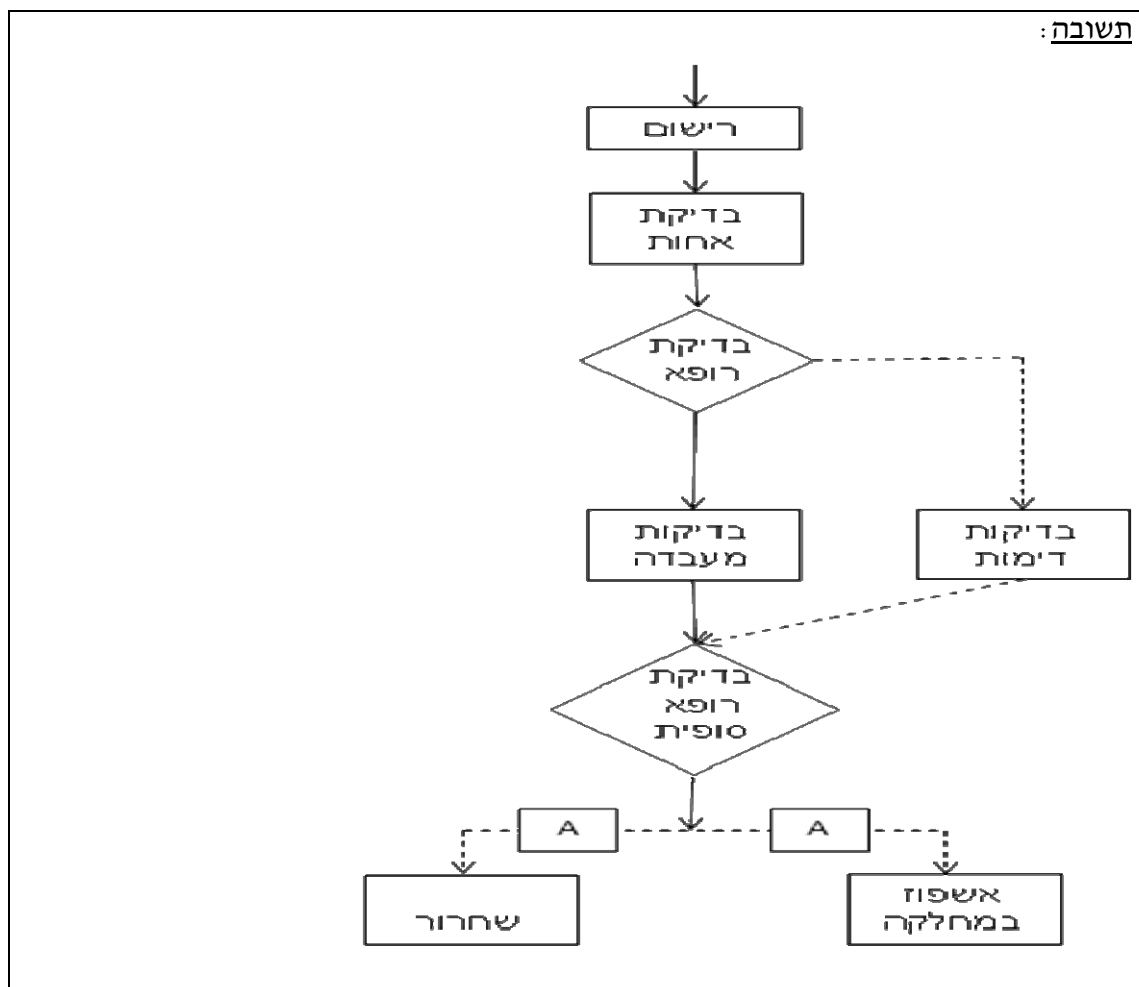
ביקור של חולה בחדר המיון מורכב, באופן כללי, מהשלבים הבאים (לפי הסדר):

- רישום בדלפק הקבלה של המיון.
- בדיקה ע"י אחות הכוללת תשאול, בדיקת סימנים חיוניים (דופק, לחץ דם), בדיקת דם (במקרה הצורך) וסווג החולה ע"פ חומרת מצבו (מ-P1 – חמור ביותר, עד P5 קל ביותר).
- בדיקת רופא.
- בדיקות מעבדה (תמיד) ולעיתים גם בדיקות דימות (נקבע בהתאם להחלטה בבדיקת הרופא).
- מפגש חוזר עם רופא לאחר קבלת תוצאות בדיקות המעבדה ובדיקות הדימות.
- שחרור או אשפוז במחלקות ביה"ח (מבחינה אדמיניסטרטיבית, השחרור או האשפוז מתבצעים ע"י פקיד/ת הקבלה של המיון)

3.1 שרטוט תרשים פעילויות וקדימויות - Activities Flow Chart של ביקור חולה בחדר

המיון.

הערה: אין צורך למדל את הפעולות המתבצעות בבדיקת אחות, מספיק להתייחס לזה כאל פעילות אחת. ניתן להתייחס לבדיקות הדימות כפעולה אחת.

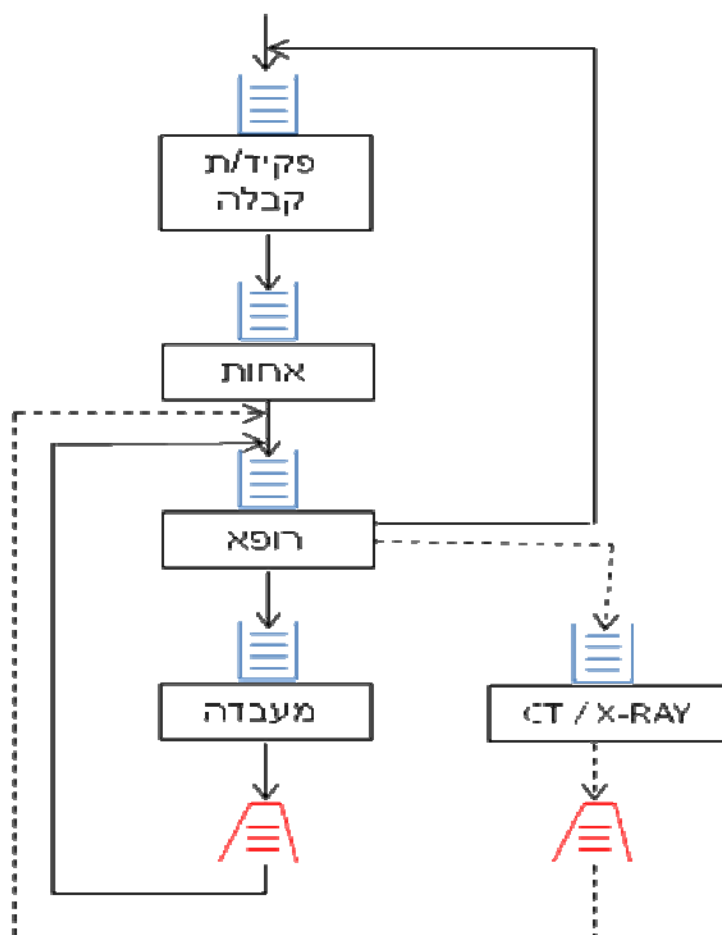


3.2. שרטטו תרשים רשת משאבים – Resources Flow Chart של ביקור חולה בחדר המיון.

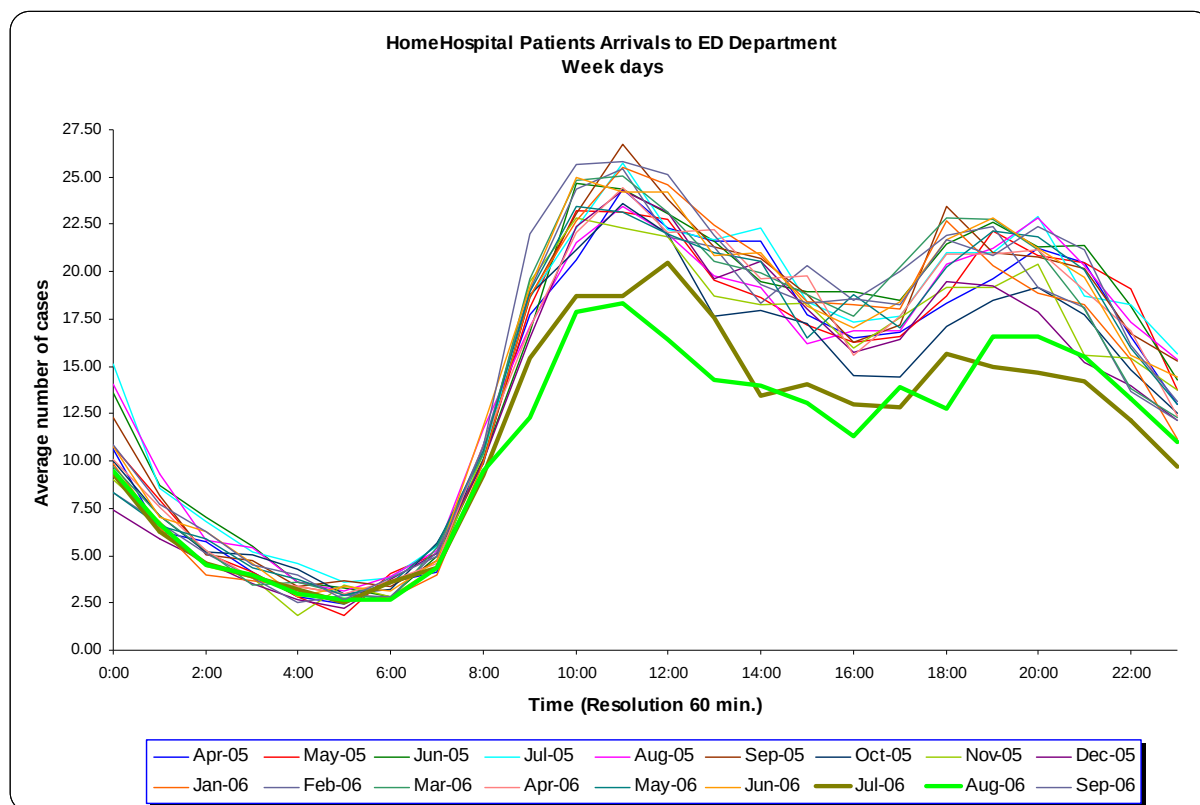
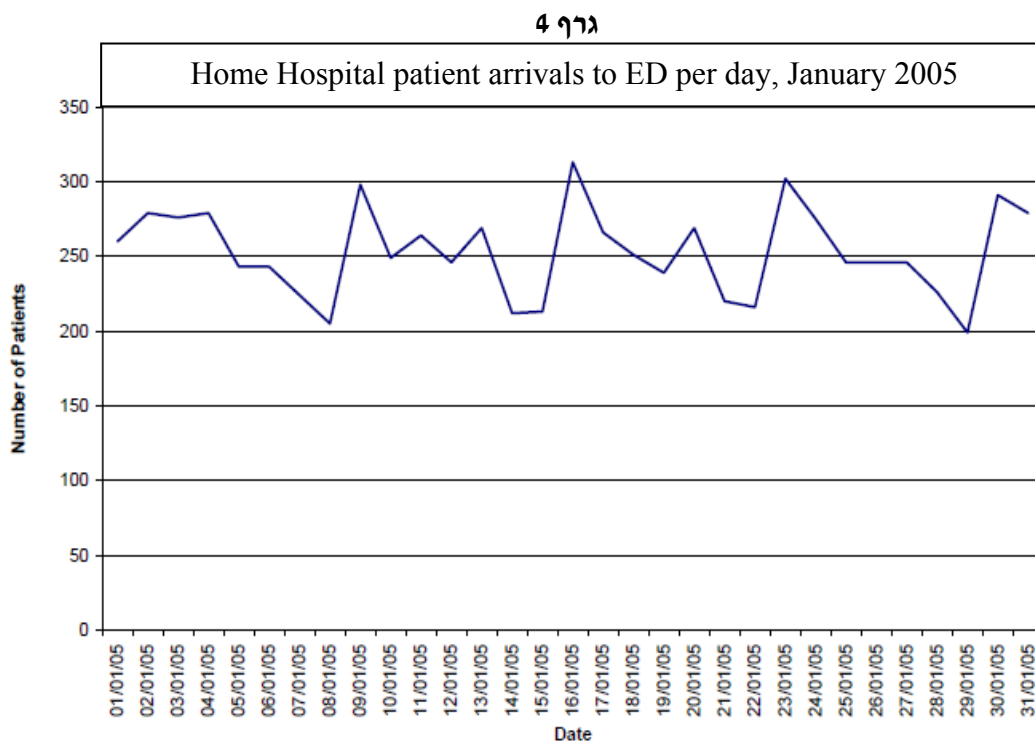
הערה:

אין צורך למדל את הפעולות המתבצעות בבדיקת אחות, מספיק להתייחס לזה כאל פעילות אחת. ניתן להתייחס לבדיקות הדימות כאל פעולה אחת (לבחור למשל, X-RAY ולהתייחס אליו בלבד).

תשובה:



גרף 4 מתאר את כמות החולים שהגיעה לחדר מיון בביה"ח בצפון הארץ, בכל יום, במהלך חודש ינואר 2005. **גרף 5** מתאר את קצב הגעת החולים לחדר מיון על פני שעות היום מאפריל 2005 עד ספטמבר 2006.



גרף 5

3.3 האם ניתן להגדיר את תהליך הגעת החולים לחדר מיון כתהליך פואסון הומוגני בזמן?

תשובה:

לא ניתן להגדיר את תהליך הגעת החולים לחדר מיון כתהליך פואסון הומוגני בזמן, קל לראות מגרף 5 כי קצב הכניסה משתנה באופן משמעותי במהלך היום.

3.4 כיצד הייתם משנים את גרף 5 ב – SEESat על מנת לבדוק האם קיימת "צורה" לתהליך הגעת החולים לחדר המיון על פני היום?

תשובה:

יש להשתמש ב – SEESat באופציה "percent to mean" אשר מציגה את כמות ההגעות בכל שעה כאחוז לעומת ממוצע ההגעות היומי, או באופציה "proportion to column totals" אשר מציגה את כמות ההגעות בכל שעה כאחוז לעומת סך ההגעות היומי.

3.5 נניח שהיו בידיכם, עבור חודש פברואר 2012, גרף דומה לגרף שיצרתם בסעיף 3.4 (צורה, נסמנו גרף A) וגרף דומה לגרף 4 (סה"כ יומי, נסמנו גרף B). כיצד הייתם משתמשים בשני הגרפים הללו על מנת לחזות את קצב הגעת החולים לחדר מיון בכל שעה ביום ראשון הראשון של חודש מרץ 2013.

תשובה:

מתוך גרף A (תוך שימוש באופציה "proportion to column totals") אנו יכולים למצוא איזה אחוז מתוך סך ההגעות היומי מהווה כמות ההגעות בכל שעה. למעשה אנו מקבלים בכל שעה את

$\lambda_{\%}(t)$ כאשר $\sum_{t=1}^{24} \lambda_{\%}(t) dt = 1$. מתוך גרף B אנו יכולים למצוא את כמות ההגעות היומית בכל

יום בחודש פברואר 2012. ע"פ שיטת הממוצע הנע (למשל) נחזה את כמות ההגעות היומית ביום

ראשון הראשון של חודש מרץ 2012 : $F_j = \frac{N_{j-7} + N_{j-14} + N_{j-21} + N_{j-28}}{4}$ כעת נציב $C = F_j$

ונחשב עבור כל שעה : $C\lambda_{\%}(t), 1 \leq t \leq 24$

כאמור, לאחר הטיפול בחדר המיון, חלק מהחולים משוחררים לביתם וחלקם מועברים לאשפוז בבית החולים. בבית החולים יש 4 מחלקות פנימיות אליהן מועברים חולים מחדר המיון. במחלקות הפנימיות יש בסה"כ 186 מיטות לאשפוז חולים. כאשר כל המיטות במחלקות הפנימיות תפוסות, חולה יועבר לאגף מיוחד במיון שם ימתין עד שתתפנה עברו מיטה באחת המחלקות הפנימיות. נמצא כי משך האשפוז הממוצע במחלקה פנימית הוא 5.12 יום וכי קצב ההגעה למחלקות הפנימיות הוא 34.4 חולים ליום. נניח כי הזמנים הבין מופעיים למחלקות הפנימיות הינם ב"ת אקספוננציאליים וכך גם משכי השהייה במחלקות הפנימיות. כמו כן נניח כי אין נטישות מהאגף המיוחד בחדר מיון, כלומר חולים אינם עוזבים אותו מרצונם ואחוז החולים הנפטרים בעת שהייה באגף זה אפסי.

3.6. איזה מודל מתמטי מתאים למידול סך מספר החולים הנמצאים במחלקות הפנימיות? מהן הנחות המודל?

תשובה:

המודל המתאים הוא Erlang B $M / M / N / N$. הנחות המודל:

- הגעות ע"פ תהליך פואסון. קצב הגעה = מופע ראשוני + מופע מהמחלקה המיוחדת (ראו הערה בהמשך).
 - משכי שירות (שהייה) אקספוננציאליים.
 - N שרתים (מיטות).
 - N מקומות במערכת.
- חולה שמגיע כאשר כל המיטות תפוסות נחסם ומועבר לאגף מיוחד במיון.
הערה: ההנחה היא שהשהייה באגף המיוחד אינה משפיעה על האופי הפואסוני של המופע הכולל לפנימיות.

3.7. נמצא כי תפוסת המיטות במחלקות הפנימיות הינה 91.7%. על סמך נתון זה ונתוני הסעיף הקודם, מהו אחוז החולים המועברים לאגף המיוחד במיון? לעזרתכם, עמודים רלוונטיים שנלקחו מההרצאות (המופיעים בשני העמודים הבאים).

תשובה:

$$\beta = \frac{n - R}{\sqrt{R}} = \frac{186 - 34.4 * 5.12}{\sqrt{34.4 * 5.12}} = 0.743$$

$$P(Block) \approx \frac{\gamma}{\sqrt{n}}$$

$$\rho \approx 1 - \frac{\beta + \gamma}{\sqrt{n}}$$

$$P(Block) \approx \frac{\gamma}{\sqrt{n}} = 1 - \frac{\beta}{\sqrt{n}} - \rho = 1 - \frac{0.743}{\sqrt{186}} - 0.917 = 0.0285 = 2.85\% \Leftarrow$$

The Erlang-A Queue in the QED-Regime

Theorem (with Garnett & Reiman, 2002)

The following **points of view** are equivalent:

- 0. **QED:** $P\{W_q > 0\} \approx \alpha,$ for some $0 < \alpha < 1$;
- 1. **Manager:** $n \approx R + \beta\sqrt{R},$ for some $-\infty < \beta < \infty$;
- 2. **Servers:** Occupancy $\approx 1 - \frac{\beta + \gamma}{\sqrt{n}};$
- 3. **Customers:** $P\{\text{Ab}\} \approx \frac{\gamma}{\sqrt{n}},$ for some $0 < \gamma < \infty$;

in which case

$$\alpha = \alpha\left(\beta, \frac{\mu}{\theta}\right) = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\hat{\beta})}{h(-\beta)}\right]^{-1},$$

which we call the **Garnett Delay-Function(s)**;

here $\hat{\beta} \triangleq \beta\sqrt{\frac{\mu}{\theta}},$ and

$$\gamma = \alpha \cdot \sqrt{\frac{\theta}{\mu}} \cdot [h(\hat{\beta}) - \hat{\beta}].$$

The Erlang-B Queue in the QED-Regime

Recall the Erlang-B Formula:

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

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

$$\gamma = h(-\beta) = \frac{\phi(-\beta)}{\bar{\Phi}(-\beta)} = \frac{\phi(\beta)}{\Phi(\beta)},$$

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.

The Erlang-C Queue in the QED-Regime

Theorem (QED Erlang-C; Halfin & Whitt, 1981)

As $n \rightarrow \infty$, the following 4 statements are equivalent:

0. **QED:** $E_{2,n} \approx \alpha$, for some $0 < \alpha < 1$;
1. **Manager:** $n \approx R + \beta\sqrt{R}$, for some $0 < \beta < \infty$;
2. **Servers:** $\rho \approx 1 - \frac{\beta}{\sqrt{n}}$;
3. **Customers:** $E[W_q | W_q > 0] \approx \frac{1}{\sqrt{n}} \cdot \frac{1}{\mu\beta}$;

in which case

$$\alpha = \alpha(\beta) = \left[1 + \frac{\beta}{h(-\beta)} \right]^{-1},$$

which we call the **Halfin-Whitt Delay-Function**.

Note: $\beta\sqrt{R}$ = **Safety-Staffing**, in analogy to Safety-Stock.

3.8 באיזה תחום תפעולי מנוהלת תפוסת המיטות? נמקו את תשובתכם.

תשובה:

תפוסת המיטות נמצאת בתחום ה-QED.

- אחוז הנחסמים עומד על כ- 2.9% (אחוזים בודדים, אך לא אפסי).
- תפוסת המיטות הינה 91.7 (ניצולת גבוהה אך לא קרובה ל-100%).
- $-1 \leq \beta \leq 1$

3.9 נתון כי משעה 16:00 עד שעה 8:00 בבוקר שלמחרת יש רופא תורן אחד בלבד בכל מחלקה פנימית. כל חולה המועבר למחלקה פנימית מסוימת חייב לעבור קבלה למחלקה ע"י רופא מחלקה זו. הזמן הממוצע של בדיקת רופא בקבלה למחלקה הוא 30 דקות. נתון כי, בממוצע, חולים ממתנים באגף המיוחד בחדר מיון עד לקבלה ע"י רופא המחלקה כשעתיים וחצי. באיזה תחום תפעולי עובדים רופאי המחלקות הפנימיות בין השעות 16:00 ל-8:00? נמקו את תשובתכם.

תשובה:

רופאי המחלקות הפנימיות עובדים בתחום ה-ED.

משך המתנה לשירות הינו פי 5 ממשך השירות וזהו מאפיין של תחום ה-ED.

3.10 מנהל המיון בוחן את הרעיון כי במקום שכל חולה יעבור קבלה במחלקה הפנימית אליה הוא מועבר, ע"י רופא המחלקה התורן, כל ארבעת הרופאים התורנים יגיעו לחדר המיון ויבצעו שם את הבדיקות ואת הליך הקבלה למחלקה. איזה מודל מתמטי מתאים לתיאור המצב כיום? איזה מודל מתמטי מתאים לתיאור המצב שבוחן מנהל המיון? מהם הפרמטרים של כל אחד מהמודלים? חוו דעתכם על הרעיון של מנהל המיון, האם שינוי זה יביא לשיפור במשך ההמתנה? מדוע?

תשובה:

המודל המתמטי המתאים לתיאור המצב כיום הוא 4 מערכות $M / G / 1$, כל אחת מתאימה למחלקה.

- תהליך כניסה פואסוני עם קצב $\frac{\lambda}{4}$, משך שירות כללי עם ממוצע $\frac{1}{\mu}$, שרת אחד.

המודל המתמטי המתאים לתיאור המערכת האלטרנטיבית הוא תור $M / G / 4$ יחיד.

- תהליך כניסה פואסוני עם קצב λ , משך שירות כללי עם ממוצע $\frac{1}{\mu}$, ארבעה שרתים במקביל

למדנו בהרצאות כי איחוד של מספר תורים עם שרת יחיד לתור אחד מרובה שרתים (Pooling)

יביא לקיצור משכי ההמתנה (בערך פי 4 במקרה הנוכחי) – "יתרון לגודל". יש לכך שתי סיבות:

1. "כולם עוזרים לכולם", אין מצב שבו שרת אחד פנוי ובתור של שרת אחר ממתנים לקוחות.
2. איחוד התורים מקטין את הסטוכסטיות היחסית במערכת.

הוחלט לבצע בביה"ח תרגיל המדמה אירוע רב נפגעים. במקרה של אירוע כזה יש להחליט באיזה סדר לטפל בנפגעים. לצורך כך, עם הגעת הנפגעים לביה"ח מתבצע תהליך הנקרא Triage. בתהליך זה מסווגים החולים לקבוצות ולכל קבוצה ניתנת עדיפות. הנפגעים בעלי העדיפות הגבוהה ביותר מטופלים קודם, אח"כ נפגעים בעלי עדיפות נמוכה יותר וכך הלאה. (למעשה גם ביום-יום, התהליך של בדיקת אחות עם הגעת החולה לחדר והחלטה על העדיפות שיש להעניק לו נקרא Triage).

לקראת התרגיל הוחלט לבחון את הכלל הטוב ביותר לקביעת סדר הטיפול בחולים. המדד המקובל באירוע רב נפגעים הוא מזעור מספר החולים הנפטרים בטרם קיבלו טיפול (נניח לפשטות כי חולים אשר מקבלים טיפול אינם נפטרים במהלכו). כל קבוצת עדיפות של חולים מאופיינת ע"י

$$\text{משך הזמן הממוצע עד לפטירה } \frac{1}{r} \text{ ומשך השירות הצפוי הממוצע } \frac{1}{\mu}.$$

נאמר שחולה A קריטי יותר מחולה B אם קצב התמותה שלו r_A גדול מקצב התמותה של חולה B

$$-r_B.$$

בדרך כלל מקובל לטפל החולים בסדר יורד של קצבי התמותה שלהם. בדוגמה הנ"ל, נטפל קודם בחולה A ואח"כ בחולה B.

3.11. הציעו כלל אלטרנטיבי לקביעת סדר הטיפול בחולים שעשוי להביא לירידה במספר החולים הנפטרים בטרם קבלו טיפול. נמקו תשובתכם.

תשובה:

הכלל שנציע הוא מתן עדיפות גבוהה יותר לחולים בעלי $r\mu$ גבוה יותר. כלומר, ניקח בחשבון גם את קצב התמותה אך גם את קצב השירות עבור כל חולה. אם כל החולים הם בעלי אותו קצב תמותה, ברור כי עדיף לטפל קודם בחולים בעלי משך השירות הקצר ביותר כי אז ממוצע משך ההמתנה עבור כלל החולים יהיה הנמוך ביותר. אם כל החולים הם בעלי אותו משך שירות, ברור כי עדיף להכניס קודם החולים בעלי קצב התמותה הגבוה ביותר. כלל זה מקביל לכלל ה- $c\mu$ שנלמד בתרגול.

דוגמה מספרית להמחשה (נניח שרת אחד):

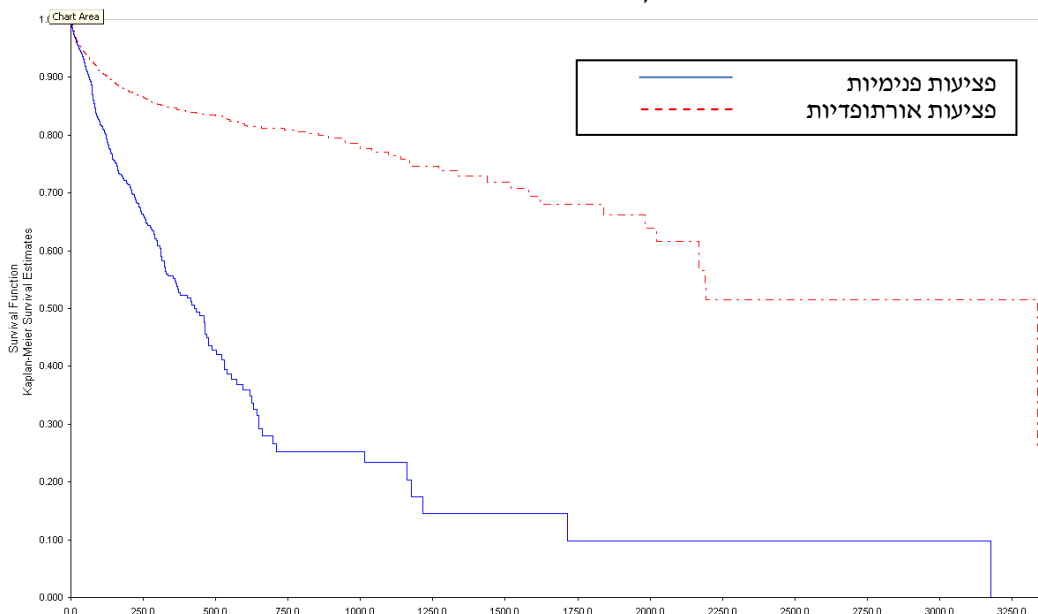
$$r_1 = \frac{1}{2}, r_2 = 1, r_3 = 3$$

$$\mu_1 = 2, \mu_2 = \frac{3}{2}, \mu_3 = \frac{1}{3}$$

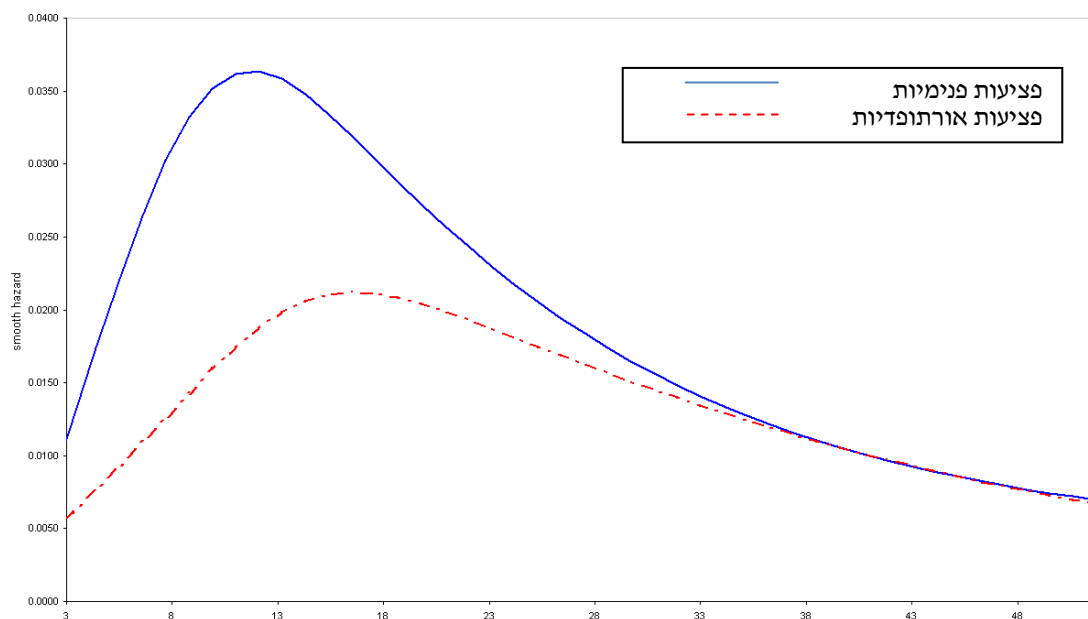
חולה מספר 3 הוא הקריטי ביותר וחולה מספר 2 הוא בעל ה- $r\mu$ הגבוה ביותר. אם נטפל קודם בחולה מספר 3, רק לאחר 3 שעות (בממוצע) נוכל לטפל בשני החולים האחרים אך סביר שהם כבר לא יהיו במערכת בשלב זה. כלומר, 2 חולים נפטרו בטרם קיבלו טיפול. אם נטפל קודם בחולה מספר 2, לאחר 40 דקות (בממוצע) נוכל לטפל בחולה מספר 1, חולה מספר 3 כבר לא יהיה במערכת בשלב זה, כלומר, רק חולה אחד נפטר בטרם קיבל טיפול.

3.12. גרף 6 מתאר את פונקציית השרידות (Survival function) של משכי השירות של שני סוגים עיקריים של פצועים המגיעים לביה"ח - פצועים הסובלים מפגיעות אורתופדיות ופצועים הסובלים מפגיעות פנימיות. גרף 7 מתאר את פונקציית הסיכון לפטירה (Hazard rate function) עבור שני סוגי המאושפזים הנ"ל. על סמך שני הגרפים הללו, לאיזה סוג לקוחות תתנו עדיפות גבוהה יותר? נמקו את תשובתכם בעמוד הבא.

גרף 6 – Survival Function



גרף 7 – Hazard rate function



תשובה:

על סמך הגרף של פונקציית השרידות, ניתן לראות כי קיים סדר סטוכסטי בין משכי השירות של שני סוגי הפצועים. ע"פ נוסחת הזנב, תוחלת משך השירות שווה לשטח מתחת לגרף של פונקציית

$$\text{השרידות} \quad E[X] = \int_0^{\infty} P(X \geq x) dx = \int_0^{\infty} S(x) dx$$

כלומר, משך השירות של פצועים עם פגיעות אורתופדיות ארוך סטוכסטית ממשך השירות של

$$\text{פצועים עם פגיעות פנימיות: } \mu_{ort} \leq \mu_{int}.$$

על סמך הגרף של פונקציית הסיכון ניתן לראות כי קיים Hazard rate order בין שני סוגי

הפצועים $X_{ort} \leq_{hz} X_{int}$, כלומר, פצועים עם פגיעות אורתופדיות הם בעלי "סבלנות" גבוהה יותר

מפצועים עם פגיעות פנימיות ולכן $r_{ort} \leq r_{int}$. לפיכך, ע"פ כלל ה- $r\mu$ שניסחנו בסעיף הקודם,

ניתן עדיפות לפצועים עם פגיעות פנימיות.

שאלה 4 – תיאוריה (10 נקודות)

4.1. מודל M/U/1 מסמן מערכת תור כאשר:

- מופעים ע"פ תהליך פואסון עם קצב λ .

- משכי שירות מפולגים Uniform $[0, b]$.

- שרת יחיד.

השוו את תוחלת זמן ההמתנה בתור זה לתוחלת זמן ההמתנה בתור שבו משכי השירות מפולגים

אקספוננציאלית עם ממוצע $b/2$ ולתור שבו משכי השירות קבועים וערכם $b/2$.

באיזה תור תוחלת זמן ההמתנה היא הארוכה ביותר? באיזה תור תוחלת זמן ההמתנה היא

הקצרה ביותר? פרטו את חישוביכם.

רמז:

$$1. \text{ עבור } X \sim \text{Uni}[0, b] : E[X] = \frac{b}{2}, \text{ Var}(X) = \frac{b^2}{12}$$

2. נוסחת Khintchine-Pollaczek לתור M/G/1:

$$E(W_q) = E(S) \cdot \frac{\rho}{1 - \rho} \cdot \frac{1 + C^2(S)}{2}.$$

תשובה:

למעשה אנו מסתכלים על שלושה תורים: M/D/1 ו M/M/1, M/U/1.

כל הפרמטרים בנוסחת Khintchine-Pollaczek לתוחלת זמן ההמתנה בשלושת התורים הנ"ל

זהים, פרט ל- $C^2(S)$.

$$\left(E[S] = \frac{b}{2}, \rho = \lambda E[S] = \lambda \frac{b}{2} \right)$$

נחשב את $C^2(S)$ ואת תוחלת זמן ההמתנה בכל אחד מהתורים הנ"ל:

$$E[W_q] = E[S] \frac{\rho}{1 - \rho} \frac{2}{3} \Leftarrow C^2(S) = \frac{\text{Var}(S)}{E(S)^2} = \frac{b^2/12}{b^2/4} = \frac{1}{3} : \text{M/U/1}$$

$$E[W_q] = E[S] \frac{\rho}{1 - \rho} 1 \Leftarrow C^2(S) = \frac{\text{Var}(S)}{E(S)^2} = \frac{1/\mu^2}{(1/\mu)^2} = 1 : \text{M/M/1}$$

$$E[W_q] = E[S] \frac{\rho}{1 - \rho} \frac{1}{2} \Leftarrow C^2(S) = \frac{\text{Var}(S)}{E(S)^2} = 0 : \text{M/D/1}$$

מסקנה: בתור M/D/1 תוחלת זמן ההמתנה היא הקצרה ביותר. בתור M/M/1 תוחלת זמן

ההמתנה היא הארוכה ביותר, והסדר נקבע ע"פ "כמות האקראיות" במשך השירות.

4.2. נניח כעת מודל M/U/1 כאשר משכי השירות מפולגים אחיד על $[0, b]$. כמו כן נניח כי, עבור כל לקוח שמגיע למערכת, אנו יודעים האם משך השירות שלו קטן מ- $b/2$ או גדול מ- $b/2$. מהי מדיניות העדיפות האופטימאלית (כלומר לאיזה סוג של לקוחות תינתן עדיפות בכניסה לשירות). הסבירו את תשובתכם.

תשובה:

ע"פ כלל ה- μc שנלמד בכיתה, ניתן עדיפות ללקוחות בעלי משכי השירות הקצרים יותר, כלומר ללקוחות שאנו יודעים שמשך השירות שלהם יהיה קטן מ- $b/2$. מדיניות זו תמזער את משך ההמתנה הכולל במערכת.

4.3. נניח מדיניות Preemptive. השוו את תוחלת זמן ההמתנה של הלקוחות בעלי העדיפות הגבוהה (על סמך תשובתכם לסעיף הקודם) לתוחלת זמן ההמתנה של לקוחות אלו בתור ללא עדיפויות. נוסחאות מהתרגול בנושא מופיעות בסוף השאלה.

תשובה:

בתור M/U/1 ללא עדיפויות:

$$E[W_q] = E[S] \frac{\rho}{1-\rho} \frac{1+C^2(S)}{2} = E[S] \frac{\rho}{1-\rho} \frac{2}{3} = \frac{\rho}{1-\rho} \frac{b}{3}$$

העדיפות הגבוהה רואה תור M/U/1 עם הפרמטרים הבאים:

- משך השירות מפולג $Uni[0, \frac{b}{2}]$.

- מופע פואסוני בקצב $\frac{\lambda}{2}$.

תחת מדיניות Preemptive, עבור העדיפות הגבוהה:

$$E(R_1) = \rho_1 E[S_1] \frac{1+C^2(S_1)}{2}$$

$$E[S_1] = \frac{b}{4}, \quad Var(S_1) = \frac{b^2}{48}, \quad C^2(S) = \frac{1}{3}, \quad \rho_1 = \lambda_1 E[S_1] = \frac{\lambda}{2} \frac{b}{4} = \frac{\lambda b}{8} = \frac{\rho}{4}$$

$$\Rightarrow E(R_1) = \rho b \frac{1}{24}$$

$$E[W_{q_1}] = \frac{E(R_1)}{1-\rho_1} = \frac{\rho}{1-\frac{\rho}{4}} \frac{b}{24}$$

$$\Rightarrow \frac{E[W_{q_without_priority}]}{E[W_{q_1}]} = \frac{1-\frac{\rho}{4}}{1-\rho} 8$$

4.4. נניח כי היינו יודעים את משכי השירות המדויקים ומשתמשים במידע זה על מנת לקבוע את העדיפויות. ברור כי ידיעה זו הייתה מאפשרת להביא לשיפור תוחלת זמן ההמתנה הכללי. נניח שהיינו רוצים לחשב את השיפור הנ"ל, כלומר, לחשב את היתרון (מהיבט משך המתנה ממוצע) שבידיעת זמני השירות המדויקים. הציעו דרך לחשב יתרון זה (אין צורך לבצע את החישוב!).

תשובה:

תארנו חלוקת $[0, b]$ ל- $(\frac{b}{2}, b]$, $[0, \frac{b}{2}]$, אינטרוולים באורך $\frac{1}{2}$. באותו אופן ניתן לחלק

לאינטרוולים באורך $\frac{1}{4}$: $(\frac{3}{4}b, b]$, $(\frac{b}{2}, \frac{3}{4}b]$, $(\frac{b}{4}, \frac{b}{2}]$, $[0, \frac{b}{4}]$, ולנתח מערכת זו כמערכת עם 4

סוגי לקוחות, עדיפויות תינתנה מהשרות הקצר (אינטרוול $[0, \frac{b}{4}]$) עד לארוך.

כך נמשיך עם אינטרוולים באורך $\frac{1}{2^n}$: $(\frac{b+2}{2^n}, \frac{b+1}{2^n}]$, $(\frac{b+1}{2^n}, \frac{b}{2^n}]$, $[0, \frac{b}{2^n}]$, עבור $n = 1, 2, 3, \dots$

הגבול יזהה עבורנו את הטוב ביותר שניתן להגיד ע"י ידע מדויק של משך השירות (אפשר להשתמש בכל חלוקן אחרת לאינטרוולים, כל עוד שאורך המקסימאלי מהם מתכנס ל-0).

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

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

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

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

The last equality can be derived via simple calculations.

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

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

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

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, $E(U) = \frac{E(R)}{1 - \rho} = E(W_q)$

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

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

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

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	$X \sim \text{Mult}(n, p)$	$p(x) = \frac{n!}{x_1! \dots x_k!} \prod_1^k p_i^{x_i}$ $x = (x_1, \dots, x_k)$	$0 \leq p_i; \sum_1^k p_i = 1; n \geq 1$ $p = (p_1, \dots, p_k)$	$0 \leq x_i; \sum_1^k x_i = n$		$E(X_i) = np_i$ $i = 1, \dots, k$	$\text{cov}(X_i, X_j) = -np_i p_j$ $(i \neq j)$	$X_i \sim \text{Bin}(n, p_i)$
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	