מעבדה 3

השערת האפס:
$H_0 : \mu_0 = \mu$

זיהוי תחנה בעלת התנהגות חריגה:

	20 ARP Messages		200 ARP Messages		1000 ARP Messages	
Measurement	8 Regular Hosts	Irregular Host	8 Regular Hosts	Irregular Host	8 Regular Hosts	Irregular Host
Count	113	20	1122	200	5987	1000
Mean	36.336	27.32	37.577	26.444	37.632	28.187
Stddev	18.055	27.633	18.733	16.883	17.996	19.305
Variance	325.98	763.594	350.946	285.035	323.878	372.685
$\bar{d}$	9.016		11.133		9.445	
df	131		1320		6985	
$\widehat{S_d}$	4.787379586		1.417383875		0.62138653	
T_{st}	1.883284966		-		-	
$T_{df,\frac{\alpha}{2}=0.05}$	1.981371815		-		-	
Z_{st}	-		8.404654668		16.5968618	
$Z_{df,\frac{\alpha}{2}=0.05}$	-		1.959963985		1.959963985	
Pass/Fail	Pass		Fail		Fail	

עבור בדיקות המדגם אל מול הייצוג הנחנו כי רמת המובהקות הנבדקת היא 5 אחוזים,
כלומר הנחת היסוד היא ש- $\alpha = 0.1$ (דו צדדי).

- **עבור 20 הודעות ARP Requests:**
  בזמן שהתחנה הבעייתית שולחת 20 הודעות ARP – זהו המדגם שלנו, שאר התחנות שולחות 113 הודעות – זו האוכלוסייה שלנו.
  הממוצע, התוחלת וסטיית התקן שונים בין האוכלוסייה לבין המדגם ומכיוון שהמדגם קטן מ-30 נצטרך להשתמש במבחן T דו צדדי.

נוסחאות:

הפרש ממוצע דגום: $\bar{d} = \mu_a - \mu_b = 36.336 - 27.32 = 9.016$

דרגות חופש: $Count_{Regular} + Count_{Irregular} - 2 = 113 + 20 - 2 = 131$

טעות תקן: $\widehat{S_d} = \sqrt{\frac{(n_a-1)\cdot S_a{}^2 \cdot (n_b-1)\cdot S_b{}^2}{n_a+n_b-2}} \cdot \sqrt{\frac{n_a+n_b}{n_a \cdot n_b}} = 4.787$

מבחן T_{st} : $T_{st} = \frac{\bar{d}-\mu_{\bar{d}}}{\widehat{S_d}} = \frac{9.016-0}{4.787} = 1.88$

לפי טבלת T: $T_{df,\frac{\alpha}{2}=0.05} = 1.981$

לפי תוצאות המבחן ובדיקה בטבלת T נקבל כי תוצאת המבחן קטנה מהתוצאה בטבלה לכן עבור רמת מובהקות של 5 אחוזים נוכל לומר כי **המדגם אכן מייצג את האוכלוסייה**, זו כמובן תוצאה מוטעית ביחס למציאות, עבור רמת מובהקות גבוהה יותר היינו יכולים לקבל את התוצאה האמיתית שמראה שהתחנה הבעייתית אכן לא מייצגת את שאר האוכלוסייה שנחשבת לתקינה.

- **עבור 200 הודעות ARP Requests:**

בזמן שהתחנה הבעייתית שולחת 200 הודעות ARP – זהו המדגם שלנו, שאר התחנות שולחות 1122 הודעות – זו האוכלוסייה שלנו.
הממוצע, התוחלת וסטיית התקן שונים בין האוכלוסייה לבין המדגם ומכיוון שהמדגם גדול מ-30 נצטרך להשתמש במבחן Z דו צדדי.

<u>נוסחאות:</u>

מבחן Z_{st} :

$$Z_{st} = \left|\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| = \left|\frac{26.444-37.577}{\frac{18.733}{\sqrt{200}}}\right| = 8.404$$

*ביצענו ערך המוחלט מכיוון שהפונקציה סימטרית.

לפי טבלת Z:

$$Z_{df,\frac{\alpha}{2}=0.05} = 1.959$$

לפי תוצאות המבחן ובדיקה בטבלת Z נקבל כי תוצאת המבחן גדולה מהתוצאה בטבלה לכן עבור רמת מובהקות של 5 אחוזים נוכל לומר כי **המדגם אינו מייצג את האוכלוסייה**, זו כמובן תוצאה שמשקפת את המציאות, במקרה הזה קיבלנו שרמת המובהקות מספיקה כדי להפריך את השערת האפס ולקבל את התוצאה שהיינו מצפים לקבל.

- **עבור 1000 הודעות ARP Requests:**

בזמן שהתחנה הבעייתית שולחת 1000 הודעות ARP – זהו המדגם שלנו, שאר התחנות שולחות 5987 הודעות – זו האוכלוסייה שלנו.
הממוצע, התוחלת וסטיית התקן שונים בין האוכלוסייה לבין המדגם ומכיוון שהמדגם גדול מ-30 נצטרך להשתמש במבחן Z דו צדדי.

<u>נוסחאות:</u>

מבחן Z_{st} :

$$Z_{st} = \left|\frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right| = \left|\frac{28.187-37.632}{\frac{17.996}{\sqrt{1000}}}\right| = 16.596$$

*ביצענו ערך המוחלט מכיוון שהפונקציה סימטרית.

לפי טבלת Z:

$$Z_{df,\frac{\alpha}{2}=0.05} = 1.959$$

לפי תוצאות המבחן ובדיקה בטבלת Z נקבל כי תוצאת המבחן גדולה מהתוצאה בטבלה לכן עבור רמת מובהקות של 5 אחוזים נוכל לומר כי **המדגם אינו מייצג את האוכלוסייה**, זו כמובן תוצאה שמשקפת את המציאות, במקרה הזה קיבלנו שרמת המובהקות מספיקה כדי להפריך את השערת האפס ולקבל את התוצאה שהיינו מצפים לקבל.

<u>לסיכום</u>: נראה כי ככל שכמות הדגימות גדלה כך נקבל שתוצאת המבחן רחוקה מהתוצאה שמתקבלת מהטבלאות, זוהי תוצאה מתבקשת שכן הוספת דגימות נותנת לנו מידע נוסף ולכן מקטינה את המרווח לשגיאות. לכן נקבל Score שגדל עבור הוספה של מספר דגימות למדגם שלנו, שכאמור אינו מייצג את האוכלוסייה.

<u>התפלגות משותפת:</u>

נראה כי התפלגות ההודעות המגיעות אל תחנה '0' היא

$$Var \sim Exp(9 \cdot \lambda) \xrightarrow{\lambda=0.5} \sim Exp(4.5)$$

מכיוון שתחנה '0' מקבלת את ההודעות מכל תשעת התחנות האחרות וכל אחת מבין התחנות שולחת הודעות מידע בהתפלגות אקספוננציאלית λ כאשר $\lambda = 0.5$, נקבל את המינימום בין ההתפלגויות ששווה לחיתוך בין המאורעות והוא מתפלג אקספוננציאלית $\tilde{\lambda} = 9 \cdot \lambda$, נראה את החישוב:

$$P(Var > a) = P(\min(Var_1 > a, Var_2 > a, \dots, Var_9 > a)) =$$

$$P(Var_1 > a) \cdot P(Var_2 > a) \cdot \dots \cdot P(Var_9 > a) = \{All\ Var_i\ are\ i.i.d\ as \sim Exp(\lambda)\} =$$

$$P(Var_1 > a) \cdot P(Var_1 > a) \cdot \dots \cdot P(Var_1 > a) = e^{-a\lambda} \cdot e^{-a\lambda} \cdot \dots \cdot e^{-a\lambda} = e^{-9\lambda a}$$

$Therefore$:

$$F_{Var}(a) = P(Var \le a) = 1 - P(Var > a) = 1 - e^{-9a\lambda} \xRightarrow{\lambda=0.5} 1 - e^{-4.5a} = 1 - e^{-\tilde{\lambda}a}$$

כעת נוכל לגזור את פונקציית ההתפלגות שקיבלנו ולקבל את פונקציית הצפיפות:

$$f_{Var}(a) = \frac{d}{da}(1 - e^{-4.5a}) = 4.5a \cdot e^{4.5 \cdot a}$$

נראה כי לפי החישוב האנליטי קיבלנו שתי פונקציות (התפלגות וצפיפות) וקל לראות שאכן פונקציות אלו הן התפלגות וצפיפות של ההתפלגות האקספוננציאלית כאשר $\tilde{\lambda} = 4.5$.

נתבונן בתוצאות המתקבלות בשני מדגמים, אחד עבור 30 הודעות ושני עבור 500 ונבצע מבחן χ^2 עבור רמת מובהקות של 5 אחוזים כדי לבדוק האם אכן האוכלוסייה הנדגמת מתפלגת $\sim Exp(4.5)$ כפי שהיינו מצפים:

- על סמך 30 דגימות:

Time Differences	Expected	Observed	Chi-sq
0 - 0.1	10.87115545	12	0.117217532
0.1 - 0.4	14.1698779	13	0.096586175
0.4 – ∞	4.533856077	5	0.047926126
		30	**0.261729833**
		$\chi^2_{df=2,\alpha=0.05}$	5.991
		Pass	

עבור המבחן נקבל:

$$\chi^2_{st} = \sum \frac{(Observed - Expected)^2}{Expected} = 0.2617$$

לפי טבלת χ^2 :

$$\chi^2 = 5.991$$

לפי תוצאות המבחן ובדיקה בטבלת χ^2 עבור רמת מובהקות של 5 אחוזים ושתי דרגות חופש נקבל כי תוצאת המבחן קטנה מהתוצאה בטבלה לכן נוכל לומר כי **המדגם מייצג את ההתפלגות המחושבת**, זו כמובן תוצאה שמשקפת את המציאות והתוצאה שהיינו מצפים לקבל.

- על סמך 500 דגימות:

Time Differences	Expected	Observed	Chi-sq
0 - 0.1	181.1859242	180	0.007762282
0.1 - 0.2	115.5292459	114	0.020242434
0.2 - 0.3	73.66469955	73	0.005997791
0.3 - 0.4	46.97068621	53	0.773942807
0.4 - 0.5	29.94983183	29	0.030123057
0.5 - 0.6	19.09685591	21	0.189662499
0.6 - 0.7	12.17669294	9	0.82874538
0.7 - 0.8	7.76420221	6	0.400866612
0.8 - 0.9	4.950673904	7	0.848316316
0.9 – ∞	6.753909664	8	0.22990256
		500	**3.33556174**
		$\chi^2_{df=9,\alpha=0.05}$	16.919
		Pass	

עבור המבחן נקבל:

$$\chi^2_{st} = \sum \frac{(Observed - Expected)^2}{Expected} = 3.335$$

לפי טבלת χ^2 :

$$\chi^2 = 16.919$$

לפי תוצאות המבחן ובדיקה בטבלת χ^2 עבור רמת מובהקות של 5 אחוזים ותשעה דרגות חופש נקבל כי תוצאת המבחן קטנה מהתוצאה בטבלה לכן נוכל לומר כי **המדגם מייצג את ההתפלגות המחושבת**, זו כמובן תוצאה שמשקפת את המציאות והתוצאה שהיינו מצפים לקבל.

ניתן לראות כי **ההפרש בין הערך בטבלה לבין ערך המבחן עלה משמעותית**, כלומר התשובה שהמבחן החזיר יותר מובהקת ב-500 דגימות מאשר ב-30 דגימות כמצופה להעלאת גודל המדגם בצורה כה משמעותית, חשוב לציין כי מספר מרווחי הזמן הוא 10 במבחן השני אל מול 3 בראשון (דרגת החופש עלתה ב-7), לכן זהו פקטור נוסף שתרם להפרש הגדול שקיבלנו יחסית לטבלה.

כעת נשנה את קצב שליחת ההודעות מתחנות 5-9 לחבילה אחת בשנייה באופן דטרמיניסטי. נתבונן בתוצאות המתקבלות בשני מדגמים, אחד עבור 30 הודעות ושני עבור 500 ונבצע מבחן χ^2 כדי לבדוק האם האוכלוסייה הנדגמת עדיין מתפלגת $\sim Exp(4.5)$:

- על סמך 30 דגימות:

Time difference	Expected	Observed	Chi-sq
0 - 0.01	1.320075545	20	264.3330368
0.01 - 0.42	24.14777019	5	15.18306247
0.42 – ∞	4.19840039	5	0.153049227
		30	**279.6691485**
		$\chi^2_{df=2,\alpha=0.05}$	5.991
		Fail	

עבור המבחן נקבל: $\chi^2_{st} = \sum \frac{(Observed-Expected)^2}{Expected} = 279.669$

לפי טבלת χ^2 : $\chi^2 = 5.991$

לפי תוצאות המבחן ובדיקה בטבלת χ^2 נקבל כי תוצאת המבחן גדולה בהרבה מהתוצאה בטבלה לכן נוכל לומר כי **המדגם אינו מייצג את ההתפלגות המחושבת**, זו כמובן תוצאה שמשקפת את המציאות והתוצאה שהיינו מצפים לקבל בהנחה שעבור חמישה תחנות קבענו שזמן שליחתן יהיה שנייה בצורה דטרמיניסטית.

- על סמך 500 דגימות:

Time difference	Expected	Observed	Chi-sq
0 - 0.1	181.1859242	342	142.7327597
0.1 - 0.2	115.5292459	40	49.37855299
0.2 - 0.3	73.66469955	23	34.84588678
0.3 - 0.4	46.97068621	24	11.23365374
0.4 - 0.5	29.94983183	19	4.003321881
0.5 - 0.6	19.09685591	9	5.338391815
0.6 - 0.7	12.17669294	11	0.113709549
0.7 - 0.8	7.76420221	6	0.400866612
0.8 - 0.9	4.950673904	5	0.000491461
0.9 – ∞	3.150058357	21	101.1474647
		500	**349.1950993**
		$\chi^2{}_{df=9,\alpha=0.05}$	16.919
		Fail	

עבור המבחן נקבל: $\chi^2{}_{st} = \sum \frac{(Observed-Expected)^2}{Expected} = 349.195$

לפי טבלת χ^2 : $\chi^2 = 16.919$

לפי תוצאות המבחן ובדיקה בטבלת χ^2 עבור רמת מובהקות של 5 אחוזים ותשעה דרגות חופש נקבל כי תוצאת המבחן גדולה בהרבה מתוצאה בטבלה, כמו במבחן הקודם שביצענו, לכן נוכל לומר כי **המדגם אינו מייצג את ההתפלגות המחושבת**.
זו כמובן תוצאה שמשקפת את המציאות והתוצאה שהיינו מצפים לקבל.

ניתן לראות כי **ההפרש בין הערך בטבלה לבין ערך המבחן נשאר די יציב**, אם היינו משתמשים בדרגות חופש זהות (נניח 2) בשני המקרים היינו מקבלים במבחן השני הפרש גדול בהרבה יחסית לראשון, כפי שציינו לעיל בזוג המבחנים הקודם העלאת דרגות החופש משפיעה על ההפרש בין תוצאת המבחן לערך בטבלה.
ההפרש שכאמור יחסית יציב בין התוצאות של שני המבחנים, מאשש את המסקנה שהעלאת מספר הדגימות ומרווחי הזמן "מקרבת" את המציאות.

<u>לסיכום</u>: כפי שציפינו לראות, כאשר משנים את תחנות 5-9 לשליחה דטרמיניסטית רואים בבירור שהמדגם החדש אינו מייצג ואילו המדגם המייצג עבור את המבחן ברמת מובהקות של 5 אחוזים כפי שנבדק, בנוסף עבור שני המקרים ראינו כי הגדלת מספר הדגימות תורמת לקבלת תשובה מובהקת יותר במבחנים שביצענו וכך גם העלאת רמות החופש במבחן.