עידן דויטש
u.multinet.co.il

סטטיסטיקה למדעי המחשב – פתרון תרגיל 4

1. ע"פ הטבלה הנורמאלית, ערכי ההסתברויות הם:

a. $\Phi(2)=0.9772$
b. $\Phi(2.2)=0.9861$
c. $\Phi(2.22)=0.9868$
d. $\Phi(-1.5)=1-\Phi(1.5)=1-0.9332=0.0668$
e. $\Phi(0)=0.5$
f. $\Phi(-5)=1-\Phi(5)=1-0.9999997=0.0000003$ (ע"פ R)
g. $\Phi(-6)=1-\Phi(6)\sim 0$

a.

i. $P(-2.22\le Z\le 1.55)=\Phi(1.55)-\Phi(-2.22)=\Phi(1.55)-[1-\Phi(2.22)]= 0.9394-[1-0.9868]=0.9394-0.0132=0.9262$

ii. $P(-1\le Z\le -0.5)=\Phi(-0.5)-\Phi(-1)=[1-\Phi(0.5)]-[1-\Phi(1)]= (1-0.6915)-(1-0.8413)=0.3085-0.1587=0.1498$

iii. $P(2\le Z\le 3)=\Phi(3)-\Phi(2)=0.9987-0.9772=0.0215$

iv. $P(Z\ge -2.5)=1-[1-\Phi(2.5)]=\Phi(2.5)=0.9938$

v. $P(Z\ge -2.33)=1-[1-\Phi(2.33)]=\Phi(2.33)=0.9901$

vi. $P(Z\le 3)=\Phi(3)=0.9987$

b.

vii. $Z_{0.75}=(0.68+0.67)/2=0.675$

viii. $Z_{0.25}=-Z_{0.75}=-0.675$

ix. $Z_{0.95}=1.645$

x. $Z_{0.975}=1.960$

xi. $Z_{0.9}=1.282$

xii. $Z_{0.025}=-Z_{0.975}=-1.960$

2. $X\sim N(10,100)$

a. $P(X\le 0)=P(X-10\le 0-10)=P[(X-10)/10\le -10/10)]=P[(X-10)/10\le -1]=\Phi(-1)$
$=1-\Phi(1)=1-0.8413=\mathbf{0.1587}$

b. ע"פ סעיף א, האחוזון הוא **0**. כלומר, $P(X\le 0)=0.1587$.

c. האחוזון ה-50 של ההתפלגות מתקבל בתוחלת, ולכן $\mathbf{Z_{50}=10}$. כמו כן, ע"פ תכונות ההתפלגות הנורמאלית, $\mathbf{Z_0=-\infty}$, $\mathbf{Z_{100}=+\infty}$.

d. $P(X\le x)=P[(X-10)/10\le (x-10)/10]=0.1 \Rightarrow (x-10)/10=Z_{10} \Rightarrow$
$x=10*Z_{10}+10 \Rightarrow x=10*(-1.282)+10 \Rightarrow x=-2.82$

3. ע"פ משפט הגבול המרכזי, סכום של משתנים מקיים בלתי תלויים נוטה להתפלגות נורמלית. גבהים של כל 2 אנשים אקראיים באוכלוסיה הוא בלתי תלוי ולכן הגיוני שהתפלגות הגבהים באוכלוסיה תהיה קרובה להתפלגות נורמלית. כמו כן, גבהים

באוכלוסיה נוטים להתפזר באופן סמטרי (יש כמות גדולה של אנשים בגובה ממוצע, ומעט אנשים גבוהים מאוד או נמוכים מאוד).

4. 3 דוגמאות לסוגי נתונים אותם לא סביר לתאר כהסתברות נורמלית:

a. הבחינה במתמטיקה בדידה באוניברסיטת ת"א מוגדרת (רשמית או לא) להיות "בחינה מסננת". לכן, באופן קבוע נכשלים בבחינה זו כ-50% מכלל התלמידים במועד א' ומעטים מאוד מקבלים ציונים גבוהים. לכן, לא סביר לתאר את ציוני התלמידים במועד א' בבחינה במתמטיקה בדידה כנתונים המתפלגים נורמאלית.

b. מספר הביקורים השנתיים בחו"ל כתלות בשכר הם נתונים שלא סביר שיתפלגו נורמאלית כיוון שהגיוני שככל ששכר של אדם עולה, הוא יוכל להרשות לעצמו יותר טיולים בחו"ל שעלותן אינה זולה. כמו כן, ייתכן שאנשים ששכרם גבוה אף יבקרו בחו"ל כחלק מהעבודה שלהם לעומת אנשים המרוויחים משכורת מינימום שהסיכוי שישהו בחו"ל כחלק מהגדרת תפקידם הוא קטן יותר.

c. מספר הזבובים כתלות בניקיון השטח הם גם נתונים שלא סביר שיתפלגו נורמאלית כיוון שיש להניח שככל שאזור מסוים מכיל זבל רב יותר (כגון מזבלה עירונית) יש סיכוי גבוה יותר שיהיו בו הרבה זבובים (הנמשכים אל הזבל כמקור אוכל). לעומת זאת, במקום שאנו רוצים לקוות שהוא נקי (כמו בית חולים למשל), מספר הזבובים צפוי להיות קטן.

5. ציונים בבחינה מתפלגים נורמלית $G\sim N(60,100)$.

a. כיוון שהתפלגות הציונים היא נורמלית והתוחלת שלה היא 60, הסיכוי לעבור את הבחינה הוא $\mathbf{P(G\geq 60) = 0.5}$.

b. **כן** (בהנחה שידוע שהתוחלת היא 60). כיוון שבהתפלגות סימטרית לאנך המקביל לציר Y, 50% משטח הגרף נמצא מימין לאנך ו-50% משטח שמתחת לגרף נמצא משמאל לאנך. לכן, הסיכוי הוא 0.5 גם במקרה זה.

c. $\mathbf{P(G\geq 80)} = P[(G-60)/10\geq(80-60)/10] = P[(G-60)/10\geq 2] = 1 - \Phi(2) = 1 - 0.9772 = \mathbf{0.0228}$

$\mathbf{P(G\geq 90)} = P[(G-60)/10\geq(90-60)/10] = P[(G-60)/10\geq 3] = 1 - \Phi(3) = 1 - 0.9987 = \mathbf{0.0013}$

d. $\mathbf{P(G\leq 20)} = P[(G-60)/10\leq(20-60)/10] = P[(G-60)/10\leq -4] = \Phi(-4) = 1 - \Phi(4) = 1 - 0.9999683 = \mathbf{0.0000317}$

e. העשירון העליון של הציונים:
$P(G\leq g) = P[(G-60)/10\geq(g-60)/10] = 0.9 \Rightarrow (g-60)/10=Z_{90} \Rightarrow g=60+10* Z_{90} \Rightarrow g=60+10*(1.282) \Rightarrow \mathbf{g=72.82}$

העשירו ן התחתון של הנתונים:
$P(G\leq g) = P[(G-60)/10\geq(g-60)/10] = 0.1 \Rightarrow (g-60)/10=Z_{10} \Rightarrow g=60+10* Z_{10} \Rightarrow g=60+10*(-1.282) \Rightarrow \mathbf{g=47.18}$

f. X – מספר התלמידים שעברו את הבחינה. זהו למעשה מ"מ בדיד בינומי. כאמור, הסיכוי לעבור את הבחינה הוא 0.5 לכן, $X\sim B(100,0.5)$ ולכן תוחלת מספר התלמידים **שעברו את הבחינה** הוא **50**=100*0.5.

נחשב את הסיכוי לקבל ציון הגבוה מ-70 בבחינה:
$P(G\geq 70) = P[(G-60)/10\geq(70-60)/10] = P[(G-60)/10\geq 1] = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$

Y – מספר התלמידים שקיבלו מעל 70. $Y \sim B(100,0.1587)$. לכן, **תוחלת** התלמידים שקיבלו מעל 70 בבחינה הוא $100*0.1587=$**15.87**. אם ברצוננו לדבר על מספר תלמידים, אז כדי לא "לחתוך" אף תלמיד, נהיה אופטימיים ונצפה שמספר התלמידים **שקיבלו מעל 70** הוא **16**.

g. נשים לב כי $H = 1.1*G \Rightarrow H \sim N(66,121)$

i. $\mathbf{P(G\geq 60)} = P[(G-66)/11\geq(60-66)/11] = P[(G-66)/11\geq -0.54] = 1 - (1 - \Phi(0.54)) = \mathbf{0.7054}$

ii. **לא.** כיוון שציון עובר הוא כבר לא תוחלת הציונים, לא ניתן לדעת מה אחוז העוברים רק מהידיעה שהתפלגות סימטרית.

iii. $\mathbf{P(G\geq 80)} = P[(G-66)/11\geq(80-66)/11] = P[(G-66)/11\geq 1.27] = 1 - \Phi(1.27) = 1 - 0.8980 = \mathbf{0.1020}$

$\mathbf{P(G\geq 90)} = P[(G-66)/11\geq(90-66)/11] = P[(G-66)/11\geq 2.18] = 1 - \Phi(2.18) = 1 - 0.9854 = \mathbf{0.0146}$

iv. $\mathbf{P(G\leq 20)} = P[(G-66)/11\leq(20-66)/11] = P[(G-66)/11\leq -4.18] = \Phi(-4.18) = 1 - \Phi(4.18) = 1 - 0.9999854 = \mathbf{0.0000146}$

v. העשירון העליון של הציונים:
$P(G\leq g) = P[(G-66)/11\geq(g-66)/11] = 0.9 \Rightarrow (g-66)/11=Z_{90} \Rightarrow g=66+11*Z_{90} \Rightarrow g=66+11*(1.282) \Rightarrow \mathbf{g=80.102}$

העשירו ן התחתון של הנתונים:
$P(G\leq g) = P[(G-66)/11\geq(g-66)/11] = 0.1 \Rightarrow (g-66)/11=Z_{10} \Rightarrow g=66+11*Z_{10} \Rightarrow g=66+11*(-1.282) \Rightarrow \mathbf{g=51.898}$

vi. X – מספר התלמידים שעברו את הבחינה. זהו למעשה מ"מ בדיד בינומי. כאמור, הסיכוי לעבור את הבחינה הוא 0.70 לכן, $X \sim B(100,0.7)$ ולכן תוחלת מספר התלמידים **שעברו את הבחינה** הוא $100*0.7=$**70**.

נחשב את הסיכוי לקבל ציון הגבוה מ-70 בבחינה:
$P(G\geq 70) = P[(G-66)/11\geq(70-66)/11] = P[(G-66)/11\geq 0.36] = 1 - \Phi(0.36) = 1 - 0.6406 = 0.3594$

Y – מספר התלמידים שקיבלו מעל 70. $Y \sim B(100, 0.3594)$. לכן, **תוחלת** התלמידים שקיבלו מעל 70 בבחינה הוא $100*0.3594=$**35.94**. אם ברצוננו לדבר על מספר תלמידים, אז כדי לא "לחתוך" אף תלמיד, נהיה אופטימיים ונצפה שמספר התלמידים **שקיבלו מעל 70** הוא **36**.

6. $T \sim N(15,36)$

a. $P(T\leq t) = P[(T-15)/6\leq(t-15)/6] = 0.3 \Rightarrow (t-15)/6 = Z_{30} \Rightarrow t = 6*Z_{30} + 15 \Rightarrow t = 6*(-0.525) + 15 = \mathbf{11.85}$

b. $\mathbf{P(T\geq 20)} = P[(T-15)/6\geq(20-15)/6] = P[(T-15)/6\geq 0.83] = 1 - \Phi(0.83) = 1 - 0.7967 = \mathbf{0.2031}$

c. **לא** צריך לדעת את תכונות ההתפלגות הנורמלית כיוון שתוחלת בכל מקרה סגורה על טרנספורמציות ליניאריות. לכן, התוחלת בפרנהייט תהיה $\mathbf{E(Tf)}=(9/5)*E(Tc) + 32 = (9/5)*15 + 32 = \mathbf{59}$

d. **לא** צריך לדעת את תכונות ההתפלגות הנורמלית כיוון שגם השונות סגורה על טרנספורמציות ליניאריות. לכן, השונות בפרנהייט תהיה
V(Tf)=V[(9/5)*Tc + 32] = 81/25*V(Tc) = 81*36/25 = **116.64**

e. **כן** צריך לדעת שההתפלגות הנורמלית סגורה על טרנספורמציות ליניאריות.
$\mathbf{Tf_{30}}$ = (9/5)*Tc_{30} + 59 = (9/5)*11.85 + 59 = **53.33**

f. Tc~N(15,36), Tf=(9/5)*Tc + 32 => Tf ~ N(59,116.64)

P(Tf≤t) = P[(Tf-59)/10.8≤(t-59)/10.8] = 0.44 => (t-59)/10.8 = Z_{44} =>
t = 10.8*Z_{44} + 59 => t = 10.8*(-0.155) + 59 = **57.326**

g. יונתן רשם בפורום להתעלם מהסעיף הזה - אז נתעלם ☺.

7. X~B(n,p)

a. n=10, p=1/10

$$P(x = n\cdot p) = p(x = 10\cdot\frac{1}{10}) = p(x=1) = \binom{10}{1}\cdot\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^9 =$$

$$\left(\frac{9}{10}\right)^9 = 0.3874205$$

b. נשים לב כי E(x)=0.1*10=1 וכי V(x)=10*0.1*0.9=0.9.

$$P(x = n\cdot p) = P(x=1) = P(x\le 2) - P(x \le 1) =$$
$$P(\frac{x-1}{0.94}\le\frac{2-1}{0.94}) - P(\frac{x-1}{0.94}\le\frac{1-1}{0.94}) = P(\frac{x-1}{0.94}\le 1.05) - P(\frac{x-1}{0.9}\le 0) =$$
$$\Phi(1.05) - \Phi(0) = 0.8531 - 0.5 = 0.3531$$

c. n=100, p=1/4

$$P(x = n\cdot p) = p(x = 100\cdot\frac{1}{4}) = p(x=25) = \binom{100}{25}\cdot\left(\frac{1}{4}\right)^{25}\left(\frac{3}{4}\right)^{75}$$
$$= 0.0917997$$

d. נשים לב כי E(x)=0.25*100=25 וכי V(x)=100*0.25*0.75=18.75.

$$P(x = n\cdot p) = P(x=25) = P(x\le 26) - P(x \le 25) =$$
$$P(\frac{x-25}{4.33}\le\frac{26-25}{4.33}) - P(\frac{x-25}{4.33}\le\frac{25-25}{4.33}) = P(\frac{x-25}{4.33}\le 1.11) - P(\frac{x-25}{4.33}\le 0) =$$
$$\Phi(0.23) - \Phi(0) = 0.5910 - 0.5 = 0.0910$$

e. נשים לב כי בסעיף d הקירוב למספר טוב יותר. הסיבה לך היא כיוון שעל פי משפט הגבול המרכזי הקירוב להתפלגות הנורמאלית גדלה ככל שגדל מספר המשתנים המקרים. אנו מבצעים בחישוב השני פי 10 יותר ניסויים (כלומר יש פי 10 יותר מ"מ בהתפלגות ברנולי) ולכן הקירוב טוב יותר.

8. $x \sim N(\mu,\sigma)$

a.

$$P(x \ge \mu+\sigma) = P(\frac{x-\mu}{\sigma} \ge 1) = 1 - P(\frac{x-\mu}{\sigma} \le 1) = 1-\phi(1) =$$
$$1-0.8413 = 0.1587$$

b. הסיכוי שהאדם הבא שנפגוש יהיה גבוה או נמוך בסטיית תקן אחת לפחות מהגובה הטיפוסי הוא פי 2 מהערך שחישבנו בסעיף הקודם. ובאופן פורמלי:

$$P(x \ge \mu+\sigma \vee x \le \mu-\sigma) = P(x \ge \mu+\sigma) + P(x \le \mu-\sigma) =$$
$$P(\frac{x-\mu}{\sigma} \ge 1) + P(\frac{x-\mu}{\sigma} \le -1) = [1-P(\frac{x-\mu}{\sigma} \le 1)] + [1-P(\frac{x-\mu}{\sigma} \le 1)]$$
$$= 2\cdot(1-\phi(1)) = 2\cdot(1-0.8413) = 2\cdot 0.1587 = 0.3174$$

c. 2 סטיות תקן:

$$P(x \ge \mu+2\sigma \vee x \le \mu-2\sigma) = P(x \ge \mu+2\sigma) + P(x \le \mu-2\sigma) =$$
$$P(\frac{x-\mu}{\sigma} \ge 2) + P(\frac{x-\mu}{\sigma} \le -2) = [1-P(\frac{x-\mu}{\sigma} \le 2)] + [1-P(\frac{x-\mu}{\sigma} \le 2)]$$
$$= 2\cdot(1-\phi(2)) = 2\cdot(1-0.9772) = 2\cdot 0.0228 = 0.0456$$

באופן דומה, הפרש של 3 סטיות תקן:

$$P(x \ge \mu+3\sigma \vee x \le \mu-3\sigma) = 2\cdot(1-\phi(3)) = 2\cdot(1-0.9987)$$
$$= 2\cdot 0.0013 = 0.0026$$

d. ראינו בסעיף ב', שהפרש של סטיית תקן אחת מהתוחלת, אינו תלוי בערך האבסולוטי של התוחלת והשונות. לכן, גם כאן הסיכוי יהיה זהה לסיכוי בסעיף ב', כלומר 0.3174.

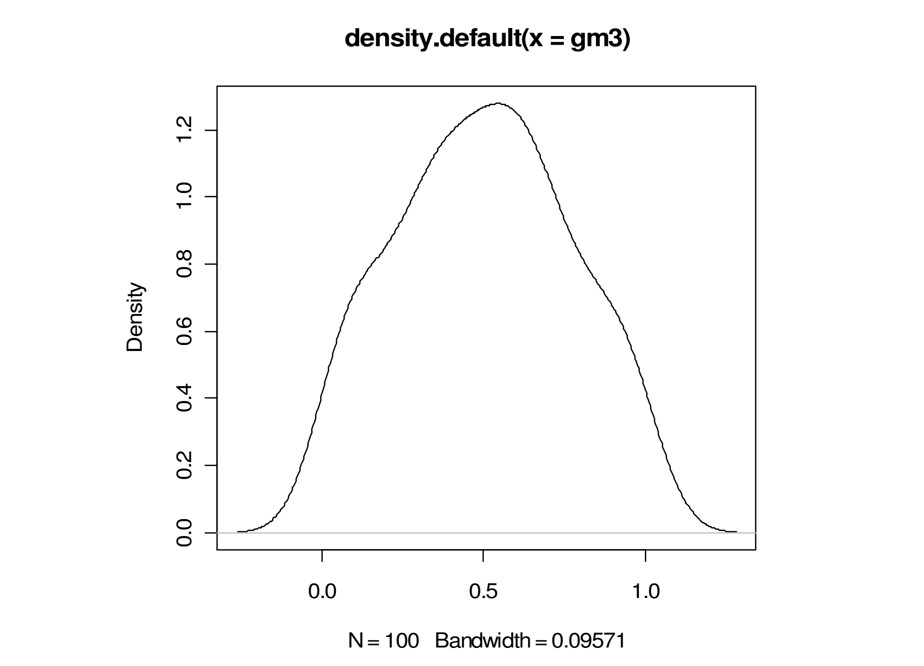
e. בסעיף ב' ראינו כי 0.3174 נמצאים בסטיית תקן אחת יותר או פחות מהתוחלת של הגבהים. כלומר, כ-69% מהאנשים נמצאים בתחום של 2 סטיות תקן (1 מעל הגובה הממוצע ו1 מתחת לגובה הממוצע). אם סטיית התקן היא 0.5 מטר, המשמעות הנגזרת מכך היא שהרוב המוחלט של האנשים נמצא בטווח של מטר בהפרש בין האנשים הגבוהים לנמוכים בקבוצה זו. מצב זה אינו אפשרי, כי בפועל, טווח הגבהים של רוב האוכלוסייה הוא (ע"פ הערכתי) כ-30 ס"מ ובטח שלא כמטר. יותר מכך, ע"פ סעיף c, 5% מהאוכלוסייה גבוהים במטר מהגובה הממוצע ו-5% מהאוכלוסייה נמוכים במטר מהגובה הממוצע. אם נעריך את הגובה הממוצע ב-1.7 מטר, הרי ש-5% מהאוכלוסייה הבוגרת צריכים להיות בגובה 70 ס"מ ו-5% נוספים בגובה 2.7 מטר. ברור שמצב זה אינו הגיוני ולא עומד במבחן המציאות.

f. (*) בהינתן תוחלת ושונות אי שוויון צ'בישב מאפשר לנו להגדיר חסמים להסתברות מסוימת. התפלגות נורמאלית גם היא דורשת מידע על תוחלת ושונות אך בניגוד לאי שוויון צ'בישב, ההתפלגות הנורמאלית נותנת חסם הדוק. כמו כן, אי שוויון צ'בישב הוא אי שוויון כללי ואינו משתמש ביתר התכונות של התפלגות נורמלית. לכן, נעדיף את ההתפלגות הנורמאלית.

9. (*)

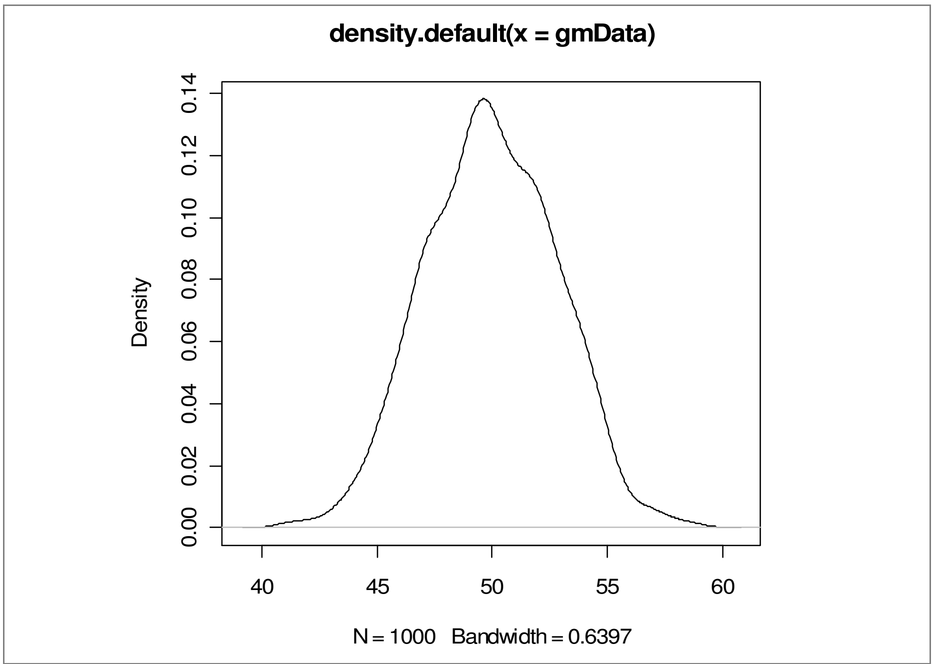
a. הגרלת המשתנים ווידוא באמצעות תרשים צפיפות:

```
> gm=runif(100,0,1)
> plot(density(gm))
```



b. ביצוע הלולאה וצפייה בנתונים באמצעות תרשים צפיפות:

```
> gmData = vector(length = 1000)
> for (i in 1:1000)
+ {
+ gmData[i]=sum(runif(100,0,1))
+ }
```



על פי התרשים נראה שהתפלגות הערכים דומה יותר להתפלגות נורמאלית.

c. אם ננסה לבדוק את 2 האפשרויות, נראה כי ככל שנוסיף סכומים, ההתפלגות תתנהג יותר ויותר כמו התפלגות נורמאלית. הסיבה לכך היא שלפי משפט הגבול המרכזי הקירוב להתפלגות הנורמאלית גדלה ככל שגדל מספר המשתנים המקרים. בדוגמה זו, כל סכום הוא מ"מ ולכן הגדלת כמות הסכומים נותנת לנו את התוצאה הרצויה.

10.

a.

i. אוכלוסיה
ii. מדגם מקרי
iii. מדגם
iv. מדגם מקרי
v. מדגם
vi. אוכלוסייה (בהנחה שהכוונה היא שאלו עוגות הכבש היחידות שאי פעם ייצרו במאפייה, אחרת, מדובר במדגם מקרי).

b. X - מספר הצימוקים בעוגה.
נחשב תוחלת ושונות:

$$E[X]=\frac{54+60+44+56}{4}=\frac{214}{4}=53.5$$

$$V(X)=\frac{(54-53.5)^2+(60-53.5)^2+(44-53.5)^2+(56-53.5)^2}{4}=$$

$$\frac{0.25+42.25+90.25+6.25}{4}=\frac{139}{4}=34.75$$