סטטיסטיקה תיאורית + לוחות שכיחות בדידים/רציפים

בגדול מקצוע הסטטיסטיקה נחלק ל-2 תחומים עיקריים- סטטיסטיקה תיאורית וסטטיסטיקה היסקית;

בסטטיסטיקה היסקית משערים השערות, משווים בין קבוצות באוכלוסיה ועוד, אך גם מסטטיסטיקה תיאורית ניתן ללמוד הרבה על האוכלוסיה- ניתן ללמוד על המבנה שלה, על הנטיות המרכזיות שלה, על הפיזור שלה וכו';

בסטטיסטיקה תיאורית עוסקים מכמה זוויות בד"כ:

1. ניתוח נתונים גולמיים.
2. ניתוח טבלאות שכיחות בדידות.
3. ניתוח טבלאות שכיחות רציפות.

הניתוחים הללו מתבטאים הן במדדים (מרכז ופיזור) והן בגרפים מתאימים (גרף עוגה/גרף מקלות/היסטוגרמה);

נתונים גולמיים הינם פשוט רצפים של מספרים שמייצגים משהו; הם יכולים להיות טבעיים, שלמים, רציונליים, מרוכבים ועוד;

טבלת שכיחות הינה למעשה דרך לקחת קבוצה גדולה של תצפיות, ולרכז אותה בטבלה, כאשר יש לנו ערכים מחד (לרוב יקראו x), ושכיחויות מנגד (frequencies).
כאשר נצפה למגוון לא גבוה של תצפיות (לדוגמא מס' ילדים במשפחה, מס' חשבונות בנק, מס' מכוניות)- לרוב ירוכזו נתונים אלה ב*טבלאות שכיחות בדידות*, וכאשר נצפה למגוון גבוה יותר של תצפיות (לדוגמא משכורת, גובה, משקל)- ירוכזו ע"פ רוב הנתונים הללו ב*טבלאות שכיחות רציפות*.

מדדי המרכז המקובלים ע"פ רוב הם ממוצע (לרוב מסומן כ- $\bar{x}$) שכיח (נקרא Mode ולרוב מסומן כ-Mo) וחציון (נקרא Median ולרוב מסומן כ-Med, Me או Md);

ממוצע- הוא סכום הערכים ברצף חלקי מספר הערכים; בכתיבה- $\bar{x} = \frac{\sum x_i}{n}$.

שכיח- הוא הערך שחוזר על עצמו הכי הרבה פעמים בהתפלגות; לעיתים מצביעים עליו ולעיתים הוא מתקבל ע"י נוסחה.

חציון- הוא הערך בהתפלגות, שמימינו ומשמאלו, ישנה בדיוק אותה כמות של תצפיות, כלומר ניתן גם להתייחס אליו לערך שלמעשה מחלק את האוכלוסיה ל-2 חלקים שווים; גם את החציון ניתן לחשב ע"פ נוסחה (נראה בהמשך).

מדדי הפיזור המקובלים ע״פ רוב הם סטיית תקן (לרוב תסומן ב- s או ב- σ), שונות (סטיית תקן בריבוע; תסומן לרוב ב- s^2 או ב- σ^2 או ב- Var - מן המילה האנגלית Variance), תחום בין-רבעוני (תסומן- IQR) ומקדם ההשתנות (לרוב יסומן כ- CV).

נרחיב על כל המדדים בהמשך;

<u>טבלאות שכיחות בדידות:</u>

כאמור בטבלאות שכיחות נמצא ערכים השייכים להתפלגות מסוימת, ולצדם- השכיחויות של אותם ערכים; לרוב לא יהיו המון ערכים כאלה;
ישנן כמה וכמה דרכים לחשב טבלאות שכיחות בדידות; אנחנו נציג דרך אחת, לא קשה במיוחד, ש׳עושה את העבודה׳.

נניח שנבדקה קבוצה של n=200 משפחות, ונתונה לנו התפלגות של מספר הילדים שם; להלן ההתפלגות שנתקבלה-

x	f(x)
0	25
1	30
2	40
3	65
4	28
5	8
6	4

נציג כעת דרך לחשב את המדדים שהזכרנו; יש להוסיף 3 טורים-

x	f(x)	x*f(x)	F(x)	x^2*f(x)
0	25	0	25	0
1	30	30	55	30
2	40	80	95	160
3	65	195	160	585
4	28	112	188	448
5	8	40	196	200
6	4	24	200	144

ממוצע- $\bar{x} = \frac{\sum x \cdot f}{n} = \frac{0+30+...+40+24}{200} = \frac{481}{200} = 2.405$ - בממוצע ישנם כ-2.4 ילדים למשפחה.

שכיח- הערך שחוזר על עצמו הכי הרבה (כלומר שה-f(x) עבורם הוא המקסימלי)- Mo=3.

חציון- הערך שמימנו ומשמאלו ישנה אותה כמות של תצפיות; כיוון שלפנינו כמות זוגית של תצפיות, החציון יהיה הממוצע של שתי התצפיות המרכזיות בהתפלגות, היינו ממוצע התצפיות במקום ה-

100 וה-101-

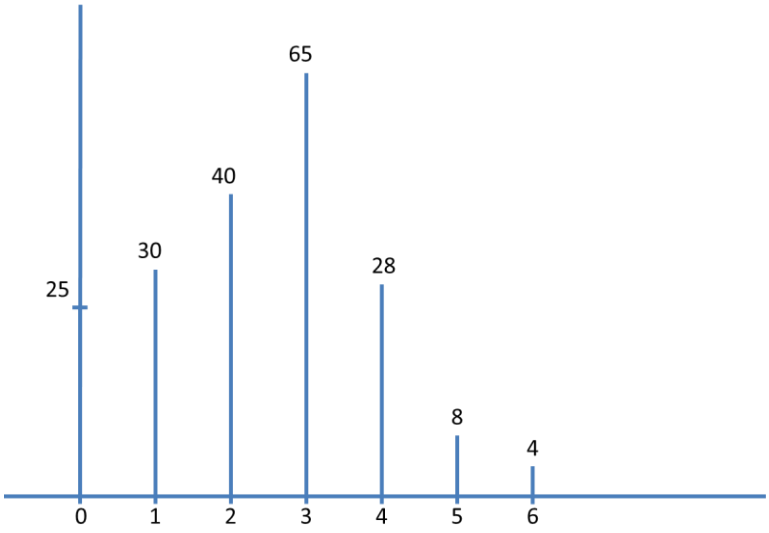
$$Med = \frac{x_{n/2} + x_{n/2+1}}{2} = \frac{x_{100} + x_{101}}{2} = \frac{3+3}{2} = 3$$

סטיית תקן- מדד המייצג את הפיזור באוכלוסיה, כלומר עד כמה האוכלוסיה הומוגנית או הטרוגנית ; סטיית התקן היא מספר טהור, 0 או יותר, שככל שהערך גבוה יותר כך סטיית התקן גבוהה יותר ; כשלעצמו אין למספר חשיבות מיוחדת, אך כאשר משווים בין אוכלוסיות, אפשר ללמוד עליהן המון מתוך ההשוואה בין סטיות התקן שלהן.
ישנן מגוון שיטות לחשב את סטיית התקן ; אנחנו נחשב את הסטייה באמצעות שימוש בנוסחת עבודה, שמקלה מאד על החישוב (יחסית לנוסחה ע"פ ההגדרה) ; נשים לב שהנוסחה מותאמת לטבלאות שכיחות, וכדי שתתאים לנתונים גולמיים יש לבצע שינויים קלים ;

כך נחשב- $\sigma = \sqrt{\frac{\sum x_i^{\,2} f(x)}{n} - \bar{x}^2} = \sqrt{\frac{1{,}567}{200} - 2.405^2} = 1.4321$.

שונות- סטיית התקן בריבוע- $\sigma^2 = \frac{1{,}567}{200} - 2.405^2 = 2.051$.

הצגה גרפית- שיטת ההצגה המקובלת במקרה של משתנה כמותי בדיד היא גרף מקלות (Bar Chart)-



<u>טבלאות שכיחות רציפות:</u>

גם כאן ישנם ערכים השייכים להתפלגות מסוימת ולצדם השכיחויות שלהם ; לרוב מדובר בערכים רבים ומגוונים, ולכן יש צורך לקבץ אותם לקבוצות ; לעיתים הקבוצות מקובצות עם גבולות מדומים ולכן יש להפוך את הקבוצות לגבולות אמיתיים ; ישנן שיטות שונות למעבר לגבולות אמיתיים ; אנחנו נתמקד בטבלאות כבר בגבולות האמיתיים שלהם ;

נניח שנבדקה אותה קבוצה של n=200 משפחות, ונתונה לנו התפלגות השכר שלהם (באלפים) ; להלן ההתפלגות שנתקבלה-

x	f(x)
0-3	15
3-5	18
5-8	55
8-12	50
12-16	37
16-20	20
20-25	5

כאן יש לבנות טבלה מעט יותר גדולה ;

x	f(x)	אמצע קטע	x*f(x)	F(x)	רוחב מחלקה	צפיפות	$x^2*f(x)$
0-3	15	1.5	23	15	3	5	33.75
3-5	18	4	72	33	2	9	288
5-8	55	6.5	357.5	88	3	18.33	2,323.75
8-12	50	10	500	138	4	12.50	5,000
12-16	37	14	518	175	4	9.25	7,252
16-20	20	18	360	195	4	5	6,480
20-25	5	22.5	113	200	5	1	2,531.25

ממוצע- מחושב באותה דרך כמו בטבלאות בדידות, רק עם אמצעי המחלקות-

$$\bar{x} = \frac{\sum \hat{x} \cdot f}{n} = \frac{1{,}942.5}{200} = 9.7125$$

שכיח- אמצע הקטע של המחלקה עם הצפיפות הגבוהה ביותר- $Mo = 6.5$.

חציון- ישנן כמה גרסאות לנוסחת החציון ; כאן תוצג נוסחה אחת, יחסית נוחה לשימוש ולעין ; את המחלקה החציונית נבחר לפי מציאת הערך n/2 (במקרה שלנו-100) ומציאת המחלקה הרלבנטית בה הוא עובר (במקרה שלנו- 8-12) ; אחרי שמצאנו את המחלקה הרלבנטית- מחלקת החציון- ***נשים לב שלמעשה החציון לא יכול להיות מחוץ לגבולות המחלקה הזו- מחלקת החציון.***

$$Med = L_0 + \frac{n/2 - F_{x(m-1)}}{f_{x(m)}} \cdot (L_1 - L_0) = 8 + \frac{100-88}{50} \cdot (12-8) = 8.96$$

ס.תקן- מחושב באותה דרך כמו בטבלאות בדידות, רק עם אמצעי המחלקות-

$$\sigma=\sqrt{\frac{\sum x_i^{\,2} f(x)}{n}-\bar{x}^2}=\sqrt{\frac{23{,}908.75}{200}-9.7125^2}=5.021$$

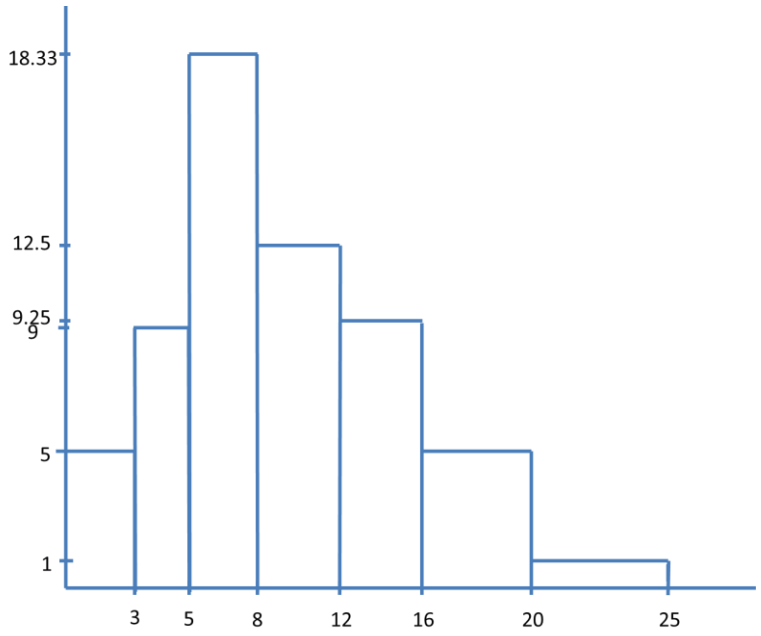
שונות- סטיית התקן בריבוע- $\sigma^2=\frac{23{,}908.75}{200}-9.7125^2=25.211$

תחום בין רבעוני- ההפרש בין האחוזון ה-75 לאחוזון ה-25 בהתפלגות, כלומר בין הרבעון השלישי לרבעון הראשון ; נחשב כ״א מהם בנפרד ולבסוף נחסיר זה מזה-

רבעון 1- $Q_1=L_0+\frac{n/4-F_{q_1-1}}{f_{q_1}}\cdot(L_1-L_0)=5+\frac{50-33}{55}\cdot(8-5)=5.927$

רבעון 3- $Q_3=L_0+\frac{3n/4-F_{q_3-1}}{f_{q_3}}\cdot(L_1-L_0)=12+\frac{150-138}{37}\cdot(16-12)=13.297$

ההפרש- $IQR=Q_3-Q_1=13.297-5.927=7.369$

מקדם ההשתנות-מדד פיזור שמשתמשים בו לעיתים ; נקרא גם CV- $CV=\frac{s}{\bar{x}}=\frac{5.021}{9.7125}=0.5169$

תצוגה גרפית- עבור משתנה רציף, ההצגה הגרפית המתאימה היא היסטוגרמה, אשר לוקחת למעשה בחשבון גם את רוחב המחלקה (ציר ה-X) וגם את הצפיפות שלה (ציר ה-Y)-