שיטות סטטיסטיות במדעי המחשב | שיעור 6: רווח סמך (confidence interval)

פרופ' רעות צרפתי | 18.2.2024

חלק ב: מושגי יסוד

נתחיל בתזכורת של מושגי היסוד שלמדנו ביחידות הקודמות:

- **אוכלוסיה** (population): אוסף של פרטים
- **משתנה** (variable): <תכונה מאפיינת> של כל פרט באוכלוסיה
- **פרמטר** (parameter): מאפיין (קבוע) של האוכלוסיה
- **מדגם** (sample): תת-קבוצה מייצגת של האוכלוסיה
- **דגימה** (sampling): השיטה לבחירת המדגם (אצלנו תמיד אקראית, רנדומית)
- **סטטיסטי** (statistic): ערך המחושב על בסיס התצפיות שבידינו

<u>בבעיית האמידה</u> יש לנו <u>מודל</u> של משתנה מקרי X~F עם פרמטר (או וקטור פרמטרי ..3 2 1)
ויש לנו מדגם מקרי X1..Xn~F . נרצה לבחור סטטיסטי במדגם כאמד לפרמטר.

- **אמד**: הגדרת הסטטיסטי באופן כללי (נוסחא)
- **אומדן**: חישוב ערך הסטטיסטי במדגם ספציפי (ערך מספרי)
- **שגיאת האמידה**: המרחק בין האמד לפרמטר בערך מוחלט | - |

אבחנה חשובה ומהותית היא שמכיוון שאין לנו גישה לפרמטר אותו רוצים לאמוד, לעולם לא נדע את ערך שגיאת האמידה! כמו כן, חשוב לזכור שהאמד (הסטטיסטי) עצמו הוא משתנה מקרי, המקבל ערכים שונים במדגמים שונים. ומכאן שגם שגיאת האמידה עצמה הינה משתנה מקרי המשתנה במדגמים השונים, ולא ניתן לחשב אותה במדויק.
מה שאנו כן יכולים לשאוף לדעת הוא, באיזה הסתברות נקבל טווח טעות מסויים מסביב לאמד הנקודתי שחישבנו. על מנת לכמת את ההסתברות לטעות בתוך טווח מוגבל כרצוננו, סטטיסטיקאים נוהגים לחשב ולדווח מדד הנקרא "רווח סמך", וזהו נושא שיעורנו.

חלק ג: רווח סמך לתוחלת במקרה של תוחלת ידועה

נתחיל לדון בנושא של חישוב רווח סמך עבור מקרה פרטי של ממוצע המדגם כאמד לתוחלת.

- נתבונן בממוצע $\bar{X}$ שהוא אמד נקודתי לתוחלת μ
- שגיאת האמידה היא $|\mu-\bar{X}|$

שימו לב: שגיאת האמידה לא ידועה לנו. שגיאת האמידה היא משתנה מקרי בעצמה.

תחת תנאים אלה, נרצה לבדוק את דיוק האמד, אך דיוק האמד לא יכול להיות מדד אבסולוטי אלא הסתברותי. **לדוגמא**: "מהי ההסתברות לכך ששגיאת הדגימה תהיה גדולה מ2?"

כמו כן, **תזכורת:**

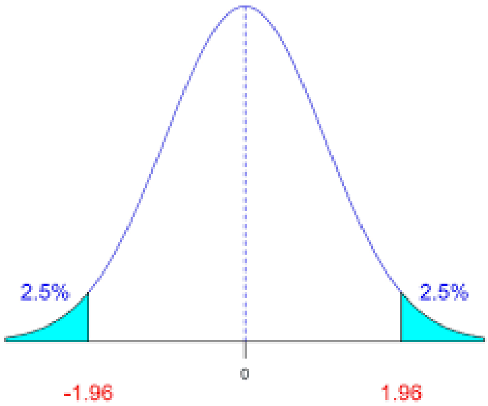
- עבור משתנה מקרי X בעל תוחלת μ, שונות σ^2, גודל מדגם n>=30
- עבור משתנה מקרי נורמלי X בעל תוחלת μ, שונות σ^2, גודל מדגם n>0 כלשהוא

מתקיים כי ממוצע המדגם מתפלג נורמלית עם הערכים הבאים:
$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$
עובדה זו מאפשרת לחשב טווח סטייה מהאמד, כמו בדוגמא הבא:

<u>דוגמא 1:</u>
במדגם שגודלו 25 מתוך מ"מ X המפולג נורמלית בעל סטיית תקן 10 ותוחלת לא ידועה, מה ההסתברות שממוצע המדגם יהיה שונה מהתוחלת בלא יותר מ 4 יחידות? נתון לנו כי
$$X \sim N(\mu, 100), n = 25$$ ומכאן אנו יכולים להסיק $$\bar{X} \sim N(\mu, 4)$$

כעת אנו יכולים לנסח את ההסתברות לקיום אי השוויון אותו מבקשים מאיתנו - ההסתברות שממוצע המדגם (הסטטיסטי) יסטה מהתוחלת האמיתית (הפרמטר) בלא יותר מ4 יחידות:
$$P(\mu-4<\bar{X}<\mu+4)$$
על ידי תיקנון (הוספת התוחלת ◌ וחלוקה בסטיית התקן של ממוצע המדגם 4) קל לראות ש:
$$P(-2<Z<+2)\bar{X}$$
ועבורה ניתן לחשב, על פי טבלת ה Z את הערכים הבאים:
$$P(-2<Z_{\bar{X}}<+2)=\Phi(2)-\Phi(-2)=2\Phi(2)-1=0.95$$

הz:



נתבונן במשמעות של תוצאה זו הנגזרת מעקומת

המשמעות עבור ההתפלגות המקורית סביב μ
היא כי השטח מתחת לעקומה
בין μ-4 ל μ+4 הוא 95%,
עם שני "זנבות" כל אחד של 2.5%.

כלומר:

אם נבצע מדגמים רבים (אינסוף), ב95% מהם
ממוצע המדגם μ יפול בטווח בין μ-4 ל μ+4

כעת נבצע את המהלך המשמעותי הבא, והוא המעבר מאי השיוויון שחישבנו עבור הסטיות סביב האמד לתוחלת, הלא הוא ממוצע המדגם לאי שיוויון דומה עבור הסטיות סביב <u>התוחלת עצמה</u>.

$$P(\mu - 4 < \bar{X} < \mu + 4)$$

$$P(\bar{X} - 4 < \mu < \bar{X} + 4)$$

כלומר, מאי השיוויון הזה:

נעביר אגפים עבור כל צד בנפרד ונקבל את הביטוי:

הביטוי התחתון נקרא: **רווח סמך ברמת סמך של 95% עבור μ**

בואו נתבונן במשמעות הביטוי השקול עבור **<u>דוגמא 1</u>:**

כלומר: במדגם שגודלו n=25 עבור מ"מ X המתפלג נורמלית עם ס"ט 10 ותוחלת μ לא ידועה, אם נבצע הרבה (אינסוף) מדגמים ונחשב עבורם רווח סמך, רווח זה "יפגוש" את μ ב95% מהמקרים.

<u>רווח סמך | המקרה הכללי:</u>

הרווח (A,B) הוא רווח סמך ברמה של 1- עבור כאשר: P(A< <B)= 1-

בואו ונראה נוסחא סגורה לחישוב רווח הסמך, כלומר לחישוב A ו B, עבור ממוצע המדגם כאמד לתוחלת, עבור משתנה מקרי X המתפלג <u>נורמלית</u> כאשר השונות שלו <u>ידועה</u>.

אנו זוכרים כי:

- עבור משתנה מקרי X בעל תוחלת μ, שונות σ^2, גודל מדגם n>=30
- עבור משתנה מקרי נורמלי X בעל תוחלת μ, שונות σ^2, גודל מדגם 0<n

כלשהוא $$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

מתקיים כי ממוצע המדגם מתפלג נורמלית עם הערכים הבאים:

מכאן ניתן לגזור את השיוויונים הבאים:

$$P(-z < Z_{\bar{X}} < +z) = \Phi(z) - \Phi(-z) = 2\Phi(z) - 1$$

$$2\Phi(z) - 1 = 1 - \alpha$$
$$\Phi(z) = 1 - \frac{\alpha}{2}$$
$$= z_{\frac{\alpha}{2}}$$

אם נשווה ל 1- את מה שקיבלנו נגיע לערך הבא:

שימו לב: "z/2" הוא nickname לשם הנוחות!

מאידך אנו יכולים לפתח את הביטוי בתוך ההסתברות לפי נוסחת התקנון ולקבל את הערך הבא:

$$P(-z < Z_{\bar{X}} < +z) = P(-z < \frac{\bar{X} - \mu}{\sigma\sqrt{n}} < z)$$
$$= P(\mu - z\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu - z\frac{\sigma}{\sqrt{n}})$$

כעת נציב את ערך ה z המתאים ל שחושב למעלה:
וכמו קודם, נעביר אגפים בשני צידי אי השיוויון ונגיע למשוואה

$$P(\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

הבאה:
ונקבל את הנוסחא הכללית הבאה:

$$P(\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

רווח סמך ברמת סמך עבור ממוצע המדגם כאשר הוא מתפלג נורמלית עם שונות ידועה הוא:

$$\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

<u>דוגמא 2 - חלק א: חישוב רווח סמך לממוצע המדגם בשונות ידועה</u>

אורך החיים של נורות מתוצרת מפעל "הניצוץ" מתפלג נורמלית עם תוחלת μ וסטיית תקן 22. נדגמו רנדומית 16 נורות ונמצא שאורך החיים הממוצע הוא 863. מצא רווח סמך ברמת 90% לμ

נתון כי אורך החיים X מתפלג נורמלית עם תוחלת μ וסטיית תקן 22. כמו כן n=16 והממוצע 863.

חשוב! נשים לב כי 1-‏z/2 =0.05 ולכן לפי ה <u>nickname</u> למעלה z/2 = 1.645 – נציב:

$$863 - 1.645\frac{22}{\sqrt{16}} < \mu < 863 + 1.645\frac{22}{\sqrt{16}}$$

<u>דוגמא 2 - חלק ב: הקשר בין רווח הסמך לגודל המדגם</u>

עבור אותה השאלה, מהו גודל המדגם שיבטיח ברמה של 90% שהערך האמיתי של התוחלת לא יהיה שונה מממוצע המדגם ביותר משעה אחת?

כאן נרצה לבטא את רווח הסמך במונחי n לא ידוע, ולחלץ אותו כדי לקבל את גודל המדגם שיבטיח קיום של רווח סמך בהסתברות נתונה 90% ובסטייה נתונה 1. נבטא את רווח הסמך ב2 דרכים :

$$\bar{X} - 1 < \mu < \bar{X} + 1$$

על פי גודל הסטייה :

$$\bar{X} - 1.645\frac{22}{\sqrt{n}} < \mu < \bar{X} + 1.645\frac{22}{\sqrt{n}}$$

על פי ערכי ההתפלגות :

קל לראות שעלינו להשוות את ערך הסטייה 1 לערך התלוי בגודל המדגם n :

$$n = (1.645 \times 22)^2 = 1309.7 \rightsquigarrow n = 1310$$

מ.ש.ל.

חלק ד : רווח סמך לתוחלת במקרה של תוחלת שאינה ידועה (מדגמים גדולים)

עד כה פיתחנו את נוסחת רווח הסמך עבור ממוצע המדגם כאשר הוא מתפלג נורמלית עם שונות ידועה. אך כפי שאמרנו מספר פעמים בעבר, ברוב המקרים השונות לא תהיה ידועה לנו, ולמעשה נצטרך גם אותה לאמוד מהמדגם. מה קורה לחישוב רווח הסמך במקרה זה? האם נשאר זהה? לא בהכרח. בחלק זה של השיעור ניגע במספר תנאים ש עבורם חישוב רווח הסמך יכול להשתנות, ונפתח גם עבורם את הנוסחא לחישוב רווח הסמך סביב ברמת סמך 1- .

נזכר כי כאשר השונות לא ידועה אנו אומדים אותה באמצעות אמד חסר הטיה :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n} \times \frac{n}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1}$$

נסמן את האמד הזה לשונות באות 2S

$$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim ?$$

שימו לב! כעת שינינו את התפלגות הדגימה!

כיצד מתפלג המ״מ ה״חדש״?

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim N(0,1)$$

<u>עבור מדגמים גדולים, כלומר n 30 מתקיים:</u>

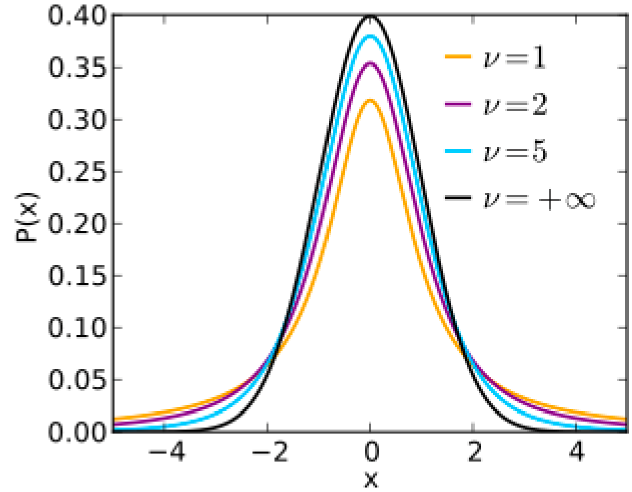
מכאן שחישוב רווח הסמך יתבצע כרגיל תוך החלפת השונות באמד חסר הטיה לשונות :

$$\bar{X} - z_{\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}}$$

חלק ד: רווח סמך לתוחלת במקרה של תוחלת שאינה ידועה (מדגמים קטנים)

מסתבר כי עבור מדגמים קטנים בהם n<30 מדובר בהתפלגות T

$$T_{\bar{X}} = \frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}} \sim t(v)$$



התפלגות T היא בעצם משפחה של התפלגויות עם פרמטר v המסמן דרגות חופש.

כש v= התפלגות T למעשה זהה להתפלגות Z. כל אמד שמחושב על סמך מדגם בגודל n מוריד לנו דרגת חופש אחת. מכאן שעבור השונות v=n-1 .

עתה כדי לחשב את רווח הסמך לא נשתמש בהתפלגות Z אלא בהתפלגות T עם ד"ח v=n-1

$$T_{\bar{X}} = \frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}} \sim t(v)$$

<u>עבור מדגמים קטנים, כלומר n< 30 מתקיים</u>

אזי חישוב רווח הסמך יתבצע כרגיל תוך החלפת השונות באמד חסר הטיה לשונות (שראינו קודם)

$$\bar{X} - t_{\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}}\frac{\hat{S}}{\sqrt{n}}$$

וכן על ידי שימוש בהתפלגות T:

Degrees of freedom (ν)	Amount of area in one tail (α)							
	0.0005	0.001	0.005	0.010	0.025	0.050	0.100	0.200
1	636.6192	318.3088	63.65674	31.82052	12.70620	6.313752	3.077684	1.376382
2	31.59905	22.32712	9.924843	6.964557	4.302653	2.919986	1.885618	1.060660
3	12.92398	10.21453	5.840909	4.540703	3.182446	2.353363	1.637744	0.978472
4	8.610302	7.173182	4.604095	3.746947	2.776445	2.131847	1.533206	0.940965
5	6.868827	5.893430	4.032143	3.364930	2.570582	2.015048	1.475884	0.919544
6	5.958816	5.207626	3.707428	3.142668	2.446912	1.943180	1.439756	0.905703
7	5.407883	4.785290	3.499483	2.997952	2.364624	1.894579	1.414924	0.896030
8	5.041305	4.500791	3.355387	2.896459	2.306004	1.859548	1.396815	0.888890
9	4.780913	4.296806	3.249836	2.821438	2.262157	1.833113	1.383029	0.883404
10	4.586894	4.143700	3.169273	2.763769	2.228139	1.812461	1.372184	0.879058
11	4.436979	4.024701	3.105807	2.718079	2.200985	1.795885	1.363430	0.875530
12	4.317791	3.929633	3.054540	2.680998	2.178813	1.782288	1.356217	0.872609
13	4.220832	3.851982	3.012276	2.650309	2.160369	1.770933	1.350171	0.870152
14	4.140454	3.787390	2.976843	2.624494	2.144787	1.761310	1.345030	0.868055
15	4.072765	3.732834	2.946713	2.602480	2.131450	1.753050	1.340606	0.866245
16	4.014996	3.686155	2.920782	2.583487	2.119905	1.745884	1.336757	0.864667
17	3.965126	3.645767	2.898231	2.566934	2.109816	1.739607	1.333379	0.863279

נבצע שימוש בטבלת ה T המתארת את ערך הזנב (Tt)Pv

שימו לב שזה בניגוד לטבלת הZ אשר מתארת את ערך ההסתברות המצטברת cdf!

זכרו גם כי ההתפלגות היא סימטרית! אזי השטח תחת העקומה בזנבות משלימים הוא שווה

[ראו דוגמת חישוב בסליידס]

חלק ה: סיכום רווח סמך עבור ממוצע המדגם

נסכם את שאמרנו על רווח סמך עבור ממוצע המדגםX כאמד ל ברמת סמך 1- :

- **ממוצע המדגם** X הוא אמד עקבי וחסר הטייה ל**תוחלת**
- חישוב **רווח סמך** עבור התוחלת ברמת סמך α ו**בשונות ידועה**:

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\bar{X} - z_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + z_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$$

- חישוב **רווח סמך** עבור התוחלת ברמת סמך α בשונות **לא** ידועה במדגמים **<u>גדולים</u>**:

$$\bar{X} - t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + t_{\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$$

- חישוב **רווח סמך** עבור התוחלת ברמת סמך α בשונות **לא** ידועה במדגמים **<u>קטנים</u>**:

חלק 1: מבט קדימה | בדיקת השערות | דוגמא ועקרונות

אנו מתמקדים פה בשיטה סטטיסטית קלאסית הנקראת "בדיקה סטטיסטית של השערת ה0"

NHST: Null Hypothesis Statistical Testing

דוגמא:

אל משרד המסחר והתעשיה הגיעה תלונה כי משקל הלחם במאפיה מסויימת נופל מהמשקל הנקוב על האריזה, 500 גרם. אנשי המשרד מודעים לכך כי לא ניתן שכל כיכר תהיה בדיוק 500 גרם, ודרישת המשרד היא כי עבור סטיית תקן המאפיינת מאפיות מודרניות (3 גרם) התוחלת של משקל הכיכר תהיה 500 גרם. המשרד החליט לבצע מבחן סטטיסטי — בדיקת השערות בשיטות סטטיסטיות — ולשם כך נדגמו 30 כיכרות לחם של אותה המאפיה, והרי משקליהם:

501 495 500 494 501 500 499 501 497 495 500 501 500 495 495 499 500 500 497 501 496 500 499 498 501 498 496 502 500 494

ממוצע המדגם הוא 498.5 גרם. האם הצרכנים צדקו בתלונתם?

כיצד נכריע מי צודק?

- המ"מ X מייצג את משקל כיכר לחם - ויכול לקבל ערכים שונים
- המ"מ $\bar{X}$ מייצג את ממוצע משקל כיכרות הלחם - ויכול לקבל ערכים שונים במדגמים שונים

ההשערה הראשונה (לטובת המאפיה) | נקראת גם השערת האפס (היא ברירת המחדל):

- תוחלת משקל כיכר הלחם היא 500 ["ובמקרה" התקבל ממוצע המדגם 498.5]

ההשערה השניה (לטובת הצרכנים) | נקראת גם ההשערה האלטרנטיבית:

- תוחלת משקל כיכר הלחם אכן קטנה מ 500 [וממוצע המדגם 500 > 498.5 מרמז על כך]

כיצד מכריעים בין ההשערות?

- מהי ההסתברות שהשערת המאפייה נכונה?
- כלומר עלינו לחשב את הערך הבא: $P(\bar{X} \leq 498.5) = ?$

אם ההסתברות שממוצע המדגם קטן או שווה לערך שקיבלנו 498.5 היא ממש ממש קטנה, כלומר קטנה מערך מובהקות מסוים , אזי יש ממש בטענת הצרכנים שככרות הלחם קטנות מידי! זאת מכיון שלא סביר שעבור השערת האפס 500 נקבל ממוצע מדגם 498.5 בהסתברות "סבירה".

ועל כן **בפועל** אנו נרצה להגדיר **אזור דחיה:**

כלומר אם הערך ההסתברותי הנ"ל קטן מערך מסוים, נניח 0.05= (שנקרא לו ערך המובהקות) אנו **נדחה את השערת האפס ונקבל את ההשערה האלטרנטיבית**.

דגים במקרה שלנו:

השערת האפס (המאפיה): היא שתוחלת משקל כיכר הלחם 500. כלומר:

$$X \sim N(500, 9)$$

$$\bar{X} \sim N(500, \frac{9}{30})$$

ואנו יודעים על פי משפט הגבול המרכזי ש :

$$P(\bar{X} \leq 498.5) \leq 0.05$$

על מנת לדחות את השערת האפס נדרש ל :

נחשב מה הערך המתקבל עבור Z בסטטיסטי שדגמנו ונשאל האם מתקיים הדבר הבא :

$$P(Z_{\bar{X}} = \frac{\bar{X} - 500}{3/\sqrt{30}} \leq -1.645) = 0.05$$

$$Z_{498.5} = \frac{498.5 - 500}{3/\sqrt{30}} = -2.73$$

אנו רואים שהערך שהתקבל הוא -2.73 > -1.645 כלומר בתוך הזנב המוגדר על ידי 0.05.

כלומר הערך נמצא באזור הדחייה, האזור שבו <u>לא סביר</u> (על פי הערך) שממוצע המדגם יימצא בו

לכן אנו **נדחה את השערת האפס** ונקבל את ההשערה האלטרנטיבית, תלונת הצרכנים מוצדקת

נסכם : שלבים בבדיקת השערות בשיטת NHST

1. **ניסוח השערות** (השערת האפס, השערה אלטרנטיבית) במונחי פרמטרים של התפלגויות
2. **בחירת הסטטיסטי** המתאים, אמד חסר הטייה לפרמטר הנבדק
3. קביעת ההנחות לגבי **התפלגות הפרמטר והתפלגות הדגימה**
4. קביעת **גודל המדגם**
5. חישוב **התפלגות הדגימה**
6. קביעת **רמת המובהקות**
7. קביעת **אזורי קבלה ודחייה**
8. **חישוב הסטטיסטי** והכרעה
