

שיטות סטטיסטיות - סיכום הרצאות לבחן

6 בינואר 2026

הסיכום נכתב תוך כדי הרצאות סמס' א' תשפ"ו (2026) ולכן ייתכן שנפלו טעויות תוך כתיבת הסיכום, ככה שהשימוש על אחוריותם.
גיא ערד-און.

תוכן עניינים

3	הרצתה 1: מבוא לקורס	1
3	שיטות מחקר:	1.1
3	מעגל החיים של ניסוי:	1.2
4	הרצתה 2: סטטיסטיקה תאורית 1	2
4	איסוף מידע	2.1
4	ממי אוספים את המידע?	2.1.1
4	מה אנחנו אוספים?	2.1.2
4	לשם מה אנחנו אוספים את המידע?	2.1.3
4	שיטות דוגמה הסתברותיות	2.1.4
5	שיטות דוגמה לא הסתברותיות	2.1.5
5	משתנים	2.2
5	סוגי משתנים	2.2.1
5	סיווג משתנים - סולמות מדידה	2.2.2
5	מדדים סטטיסטיים	2.2.3
6	תיאור והצגה	2.3
9	הרצתה 3: סטטיסטיקה תאורית 2	3
9	מדדים סטטיסטיים	3.1
10	חישוב מדדי מיקום מרכז עבור מחלקות עם גבולות אמיתיים	3.2
11	מהו המדד הכי טוב עבור \bar{x} מיקום מרכז?	3.2.1
12	מדדי פיזור	3.3
12	שונות וסטיתת תקן	3.4
13	מוצע משוקלל ושונות מצורפת	3.5
14	מדדי קשר בין מספר משתנים	3.6
15	הרצתה 4: הסקה סטטיסטית	4
15	מבוא	4.1
15	הסקה סטטיסטית	4.2
15	מושגים בסיסיים	4.3
16	התפלגות דוגמה	4.4

18	הרצאה 5 : אמידה סטטיסטית נקודתית	5
18	אמידה סטטיסטית	5.1
18	בעיית האמידה	5.2
19	תכונות של אמדים	5.3
22	שיטות אמידה	5.4
22	שיטת המומנטים	5.4.1
23	שיטת הניראות המרבית	5.4.2
25	הרצאה 6 : אמידה סטטיסטית של מרוחבי בטחון	6
25	רוחם סמך של ממוצע המדגם	6.1
26	רוחם סמך - הגדרה פורמלית	6.2
26	רוחם סמך לממוצע עבור התפלגות נורמלית כאשר השונות ידועה כאמור	6.3
26	לתוכלת	
27	רוחם סמך לממוצע כאמור לתוכלת כאשר השונות אינה ידועה	6.4
29	הרצאה 8 + 7: הסקט סטטיסטית - בדיקת השערות	7
29	השערת האפס וההשערה האלטרנטטיבית	7.1
30	סוגי השגיאות	7.2
30	חישוב הסבירות של השערת האפס	7.3
31	שלבים להחלטה אם לקבל או לדחות את השערת האפס	7.4
31	תיקון ל מבחנים מרובים	7.5
32	חישוב גודל המדגם הדרושים	7.6
32	מספר מוגדים	7.7
33	גודל האפקט (d של כהן)	7.8
33	סיכום חשוב	7.9
33	מבחני השערות במוגדים גדולים	7.10
34	מבחני השערות	7.11
34	דוגמה ראשונה: ממוצעים	7.11.1
35	דוגמה שנייה: הצלחות במוגדים	7.11.2
35	דוגמה שלישיית: הפרשים בין ממוצעים	7.11.3
37	מבחנים של זוגות	7.11.4
38	מבחנים אפרמטריים: מבחן χ^2	7.11.5
38	מדידת הפרשים א-פרמטרית במוגדים בלתי תלויים:	7.11.6
38	מבחן Mann Whitney U TEST	
38	מדידת הפרשים א-פרמטרית בדגם מזוג (מבחן Test Wilcoxon Singed-Rank	7.11.7
41	(Wilcoxon Singed-Rank	
42	סיכום מבחני השערות	7.12
43	הרצאה 9: (Analysis of Variance) Anova	8
44	הנחות יסוד למבחן אנוובה (ביסי)	8.1
44	הרעיוון המרכאי של מבחן אנוובה	8.2
44	הגדרה פורמלית	8.3
48	כיצד מתגברים על ההנחהות?	8.4
48	איך מתגברים על ההנחה שהנתונים מתפלגים נורמלית?	8.4.1
48	כיצד נדע איזו קבוצה שונה?	8.5
49	הפרש הממוצעים המינימלי שהוא כוביוק סטטיסטית	8.5.1
49	קירוב של q	8.5.2
49	סיכום HSD	8.5.3
50	אנוובה בשתי משתנים	8.6

51	הרצתה 10: רגرسיה לינארית	9
52	הגדרה פורמלית - מציאת קו המגמה	9.1
53	שיטה שנייה (שיטת ההסתברותית)	9.2
54	שיטה שלישית (אלגברה לינארית)	9.3
56	האם הקשר לינארי?	9.4
56	חישוב מובהקות סטטיסטיות	9.5

1 הרצתה 1: מבוא לקורס

סטטיסטיקה: תחום ידע שנוגע לאיסוף, עיבוד ניתוח והסקת מסקנות מנתונים ממשותיים. מחלקים את הסטטיסטיקה לשני תחומי דעת: תאוריית, והיסkit.

סטטיסטיקה תאוריית: עוסקת בתיאור תמציתי וקל לתפיסה של אוכלוסייה על סמך מדדים. למשל: יציג ע"י דיאגרמה, מדדי מקום כמו ממוצע שכיח וחיצון, מדדי פיזור כמו שונות וסטיית תקן.

סטטיסטיקה היסקיט: עוסקת בניסיון להגעה למסקנות לגבי אוכלוסייה על סמך מדגם. (למשל: סקר בחרות)

אמידה סטטיסטית: אלו שיטות מתמטיות שמאפשרות לגזר מותוך נתוני המדגם אומדן ערך של משתנה עבור אוכלוסייה. הבסיס למידה חישובית.

בדיקה השערות: כלים מתמטיים לבחינת תפקות תוצאות ניסויים לגבי משתנה או קשר בין משתנים. הבסיס לחקירה מדעית.

אמפירי = מבוסס על ניסוי

1.1 שיטות מחקר:

כיצד אנו רוכשים ידע על העולם?

הגשה הרציונלית: על ידי היקשים והסקת מסקנות. (למשל: אם כל האנשים בני תמותה, ורעות היא בת אדם, גם רעות היא בת תמותה).

הגשה האמפירית: ידע מבוסס על תצפית נסיוון ומדידה. (למשל: השימוש זרחה הבוקר, היא תזרח גם אחר).

הגשה המדעית = הגשה האמפירית + הגשה הרציונלית.

מטרת הגשה המדעית: להבין עבר, לנבأ עתיד ובעיקר לנ Sach תאוריות.

תאוריה מדעית: מערכת מונחים, הגדרות וطنאות. התאוריה כוללת מערכת של טענות על קשרים בין מונחים.

ניסוח בעית מחקר: בעית מחקר היא בעיה שנותן לחוקר אותה בכלים מדעיים. הבעיה צריכה להיות מנוסחת בצורה אובייקטיבית, ברורה וחד משמעית. הבעיה צריכה לבטא יחס בין שניים או יותר משתנים. הבעיה חייבת לעמוד בביקורת אמפירית (דרך למדידת משתנים)

השערת מחקר - מומקדת וספציפית, משקפת את ציפיות החוקר וכן יש את קוריטריון ההפרכה: השערה שיש דרך אמפירית להפריך אותה - מערכ ניסוי.

1.2 מעגל החיים של ניסוי:

איסוף מידע (הרצתה 2) ← → תיאור והצגה (הרצתה 3) ← → אומדן פרמטרים (הרצתה 6 – 4) ← → בדיקת השערות (הרצתה 9 – 7) ← → ניסוח השערה ← → (וחזר על עצמו)

אומדן מידע + תיאור והצגה = סטטיסטיקה תאורית.
אומדן פרמטרים + בדיקת השערות = סטטיסטיקה היסקית.

2 הרצאה 2: סטטיסטיקה תאורית 1

2.1 איסוף מידע

2.1.1 מיי אוסף את המידע?

א. **אוכלוסייה** - אוסף של אנשים, דברים, האובייקטים אותם אנו רוצים לחקור.

- ב. **מדגם** - תת קבוצה (מייצגת) של האוכלוסייה
1. אם קיים קושי במידה של האוכלוסייה כולה (מסובכת, ארכובה, יקרה)
2. קושי באיסוף המידע (הרבה מידע)
3. עצם המדידה פוגע בתוכונה (כלומר, אם למשל אנחנו במבצע גפרורים ורוצים לבדוק את מס' הגפרורים התקינים - שביל לבדוק אם הוא תקין נטרך להשתמש בו ולכן הפכנו אותו ללא תקין. אם נkeh את כל הגפרורים ונבדוק אותם נשאר ללא גפרורים תקנים: لكن אנחנו חיברים לחתת מדגם).

מה הכוונה בתת קבוצה מייצגת? קבוצה ששמורה את התכונות של האוכלוסייה, משמרת את הפייזו וניתן להכליל ממנה.

ג. **דגימה** - שיטת הדגימה של תת קבוצה מייצגת (השיטה בו אנו בוחרים את המדגם).

2.1.2 מה אנחנו אוספים?

משתנה: תוכנה הניתנת לתצפית ומדידה עברו כל אלמנט באוכלוסייה.

ערך: הערך שנמדד עברו אלמנט אחד באוכלוסייה.

מידע: הערכים שנמדדעו עברו כל האוכלוסייה.

2.1.3 לשם מה אנחנו אוספים את המידע?

סטטיטיסטי: ערך המוחשב על סמך הדadata, כלומר על סך כל הערכים שנמדדעו. (ממוצע הדגימות).

פרמטר: מאפיין של האוכלוסייה. למשל, תוחלת ההתפלגות.

—♥— המשתנה הוא תוכנה, למשל אם נבצע מדגם אודוטס סכום הכסף הממוצע שסטודנט מוציא בשנה א', וקיים שמהמוצע הוא \$ 178, אז הסטטיסטי הוא \$ 178. וכן המשתנה הוא סכום הכסף הממוצע.

—♥— **יתכן סטטיסטים שונים:** למשל מינימום מקסימום, חציון, שוננות וכו'.

2.1.4 שיטות דגימה הסתברותיות

בשיטות דגימה הסתברותיות ישנה הסתברות שווה לכל פרט להבחר.

1. דגימה אקראית - רנדומית. דגימה של k איברים מתוך N . זו דגימה שיכולה להתבצע עם החזרה או ללא החזרה. **ב庫ס זה כאשר נאמר כי אין מודדים** - **נמדדוד לפי דגימה אקראית**.
2. דגימה בשכבות - חלוקת האוכלוסייה לשכבות גזות ומשלימות. דגימה רנדומית (לפי פורפראציה) מכל שכבה. דוגמה לדגימה בשכבות: סקרי בחירות. לדוגמה שכבות אוכלוסייה - אם ידועים שישנם 27% מהאוכלוסייה בגילאים 50 – 40 אז דוגמים פורפראציונלית משכבות גיל.

3. דגימת אשכולות - חלוקת כל האוכלוסייה לקבוצות זרות ומשילומות. דגימה רנדומית של קבוצות והוספת כל הפרטים מכל קבוצה למדגם. למשל: ביצוע סקר מודד האושר. במקום מודד אחד אחד, אפשר למודד בתים אב. אם בית אב יצא $5/5$ במדד האושר - כל האנשים בבית האב ה"ל ייחסו $5/5$ במדד האושר. לעומת זאת כולם ייחסו את כולם.

2.1.5 שיטות דגימה לא הסתברותיות

ישנן שיטות דגימה שאינן הסתברותיות.

1. **דגימת נוחות** - "מן המוכן", הכל בביטחון אחד. כלומר מקבלים את הדגימה בביטחון אחד. למשל: משאל רחוב, מקבלים את התוצאות מיד. מה טוב בשיטה? מהיר. מה עייתי? לא מייצג את האוכלוסייה.

2. **דגימת שיפוטית** - לפי שיקול דעת החוקרת, לפי מענה על שאלונים. מה טוב בשיטה זו? אנחנו מניחים שהחוקרת יודעת מה היא עשויה ולפניהם זה טוב לנו שהיא בוחרת את האוכלוסייה. מה לא טוב? לא מייצג וסיכוי גבוה להטיהה.

3. **דגימת בדור שלג** - "חבר מביא חבר". כלומר - ניסויים שאים מגע, מקבל כסף על הניסוי ואומרים לו להביא חברים לניסויו שיבוא גם הוא "להרוויח כסף". יתרון: קל וממיר, דגימת אוכלוסייה זהה. חסרון: לא מייצג, ישנה הטיהה, מדגם של חלק ספציפי באוכלוסייה.

ישנן **סכנות לתקיפות הניסוי**: דגימה לא מייצגת / מוטה, דגימה "התנדבותית", דגימה קטנה מדי. למה להשתמש בשיטות דגימה לא הסתברותיות? פילוט, מעוט אקוטי, תוצאות מאוד נדרות.

בקורס זה נשתמש בשיטות דגימה הסתברותיות.

2.2 משתנים

2.2.1 סוגי משתנים

א. **קטגוריאי**: קבוצות ערכיים סופיים. קטגוריה מדרנית, דרגה בצבאי, קבוצת המדינות = $\{XS, S, M, L, XL, XXL\}$.

ב. **מספריים**: מס' הסטודנטים בקורס, מספר אסיטים לשחקן, גובה משקל וכו'.

המשתנה הבודד: קבוצת ערכים סופית ובת מניה.

המשתנה הרציף: קבוצת ערכים אינסופית, בין שני ערכים קיימים ערך. למשל - מרחק.

2.2.2 סיווג משתנים - סולמות מדידה

סולם שמי: יחס זהות, ללא יחסי. למשל: קטגוריה מגדרית, אرض לידה. - משתנה קטגוריאי. כלומר, אין יחס סדר מי גדול יותר אלא רק יחס שייכות.

סולם סדר: יחס זהות, עם יחס סדר. למשל: תווית מדידה, דרגה אקדמית. - משתנה קטגוריאי. בסולם זה כן יש יחס זהות, כל אחד משתייך לדבר מסוים אך יש יחס סדר בין הדברים.

סולם רוחתי: עם יחס סדר, עם מרווחים קבועים. למשל: טמפרטורה - משתנה מספרי. בסולם זה: יש משמעות למרווחים בין הערכאים. למשל: מרווחים לmorphisms בין הטמפרטורות השונות.

סולם מנה: יחס סדר, מרווחים קבועים, נקודת אפס. למשל: גובה, משקל. - משתנה מספרי. 0 מייצג העדר תוכנה.

2.2.3 מדדים סטטיסטיים

סטטיסטי: ערך המוחשב על סמך התוצאות בפועל. למשל - ממוצע, חציון. פרמטר: תוכונה של האוכלוסייה המקורית. למשל - תוחלת בהתפלגות נורמלית, פורפרציה בהתפלגות בינומית.

– בחלק של "סטטיסטיקה תאורית", נשתמש בסטטיסטיים לתיאור הדאטא. בחלק "הסקה סטטיסטית" – נשתמש בסטטיסטיים כאומדן לפרמטרים.

2.3 תיאור והצגה

כיצד ניתן להציג את המידע שנאסף?

* **תצוגה טבלאית**

1. טבלת שכיחויות. למשל פונקציה $N \rightarrow \{0, 1, \dots, 100\}$: f שמקבלת ציון v ו(v) $f(v)$ זה מס' הסטודנטים שקיבלו אותו.
2. שכיחות יחסית: $f(v) / N$. בטבלה מטה סה"כ שכיחות שמסתכמת ל-20. שכיחות יחסית תהיה האחוז של הערך v

ערך v	שיעור $f(v)$	שיעור יחסית $f(v) / N$
2	3	3/20 = 0.15
3	5	5/20 = 0.25
4	3	3/20 = 0.15
5	6	6/20 = 0.30
6	2	2/20 = 0.10
7	1	1/20 = 0.05

3. שכיחות יחסית מצטברת: RF . כמו יחסית, רק כל ערך צובר את השכיחות של הערך הקודם:

ערך v	שיעור $f(v)$	שיעור יחסית $f(v) / N$	שיעור יחסית מצטברת $RF(v)$
2	3	3/20 = 0.15	0.15
3	5	5/20 = 0.25	0.15+0.25 = 0.4
4	3	3/20 = 0.15	0.4+0.15 = 0.55
5	6	6/20 = 0.30	0.55+0.30 = 0.85
6	2	2/20 = 0.10	0.85+0.10 = 0.95
7	1	1/20 = 0.05	0.95+0.05 = 1.00

4. משתנה מספרי בדיאן: חלוקה למחלקות

ניתן לחלק את הערכים השונים למחלקות. למשל במקומות להציג 10, 1, 2, 3, 4 – 7, 8 – 10 במחלקות שונות. לשם כך צריך לדאוג שהמחלקות יהיו זרות, חלוקה ממזכה שאיחודם הוא כל ערכי המדגם ושמירה על גבולות דמיוניים בין המחלקות.

5. משתנה מספרי רציף: חלוקה למחלקות

הגבול העליון של מחלוקת אחת מתלכד עם הגבול התחתיו של זו שאחריה. בהינתן מחלוקת $[x_0, x_k]$

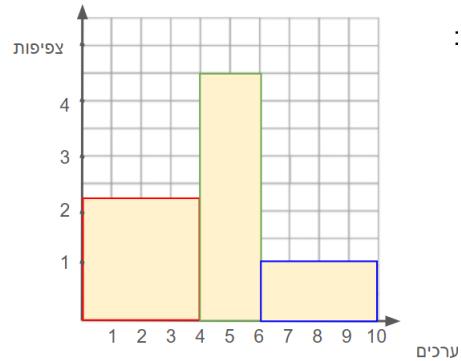
רוחב מחלוקת: הפרש בין גבול תחתיו אמיתי לגבול תחתון אמיתי. יתקיים $I = x_k - x_0$

מחלקה פתוחה: רק גבול עליון או תחתון

מחלקה	שיעור f	שיעור מחלוקת 1	רווח טווח	מרכז טווח	כפיפות $d=f/I$
0-4	9	4-0 = 4	0 + 4/2 = 2	9/4 = 2.25	
4-6	9	6-4 = 2	4 + 2/2 = 5	9/2 = 4.5	
6-10	4	10-6 = 4	6 + 4/2 = 8	4/4 = 1.0	

$$\text{מרכז טווח של מחלוקת } x_0 + \frac{I}{2} : [x_0, x_k] \quad d = \frac{f}{I}$$

מכאן מקבילים היסטוגרמה: גרען שמייצג את הערכים. השטח מתחת להיסטוגרמה הוא סה"כ השכיחויות.



כיצד בונים היסטוגרמה?

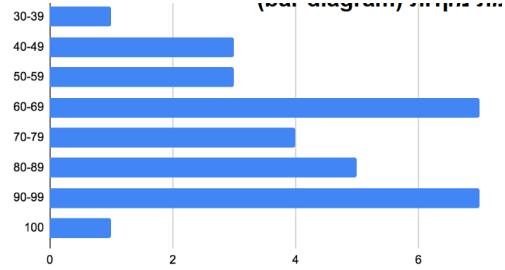
- מחליטים על מס' המחלקות שנרצה - k
 - מחשבים את הטווח של ההיסטוגרמה $r = \max - \min + 2$ (מוסיפים פלוס 2 רק באשר אנחנו יודעים את הערכים עצם ממש ולא ההיסטוגרמה).
 - מחשבים רוחב כל מחלקה $a = \frac{r}{k}$
 - מחשבים גבולות מדומים - $\min - a$
 - בחירה ייחודית הדיק u
 - чисוב גבולות אמיטיים $u - \min$
- היסטוגרמה נcona רק כאשר נתונים לנו כל הנתונים. אם נתון לנו טבלת שכיחיות אי אפשר לעשות זאת.

* תצוגה גרפית:
מייצגים נתונים באמצעות דיאגרמה.

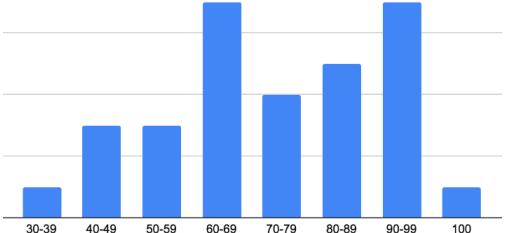
- דיאגרמת גבעול-עליה:** דיאגרמה בה מפצלים את הערכים לעשרות ויחידות. חלוקת טווח הערכים לנבעולים. וכן פירוט ערכי העליים.

stem	leaf	האחדות
3	3	
4	2 9 9	
5	3 5 5	
6	1 3 7 8 8 9 9	
7	2 3 4 8	
8	0 3 8 8 8	
9	0 2 4 4 4 4 6	
10	0	

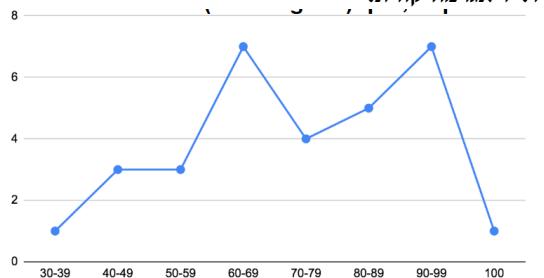
ב. **דיאגרמת מקלות:**



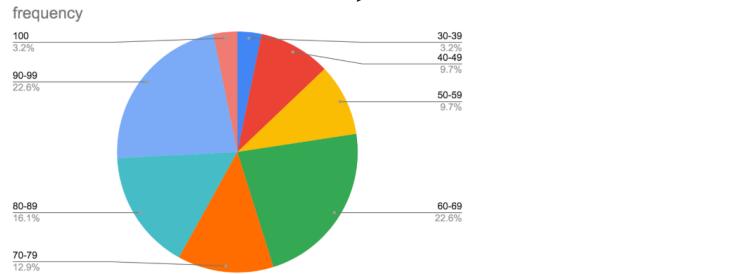
ג. דיאגרמת עמודות:



ד. דיאגרמה קוית:



ה. דיאגרמת עוגה:



מתי השתמש באיזו דיאגרמה?

עבור כל סולמות המדידה: טבלת שכיחיות, דיאגרמת עמודות, תרשימים עוגה. עבור סולם רוחים או סולם מנה: היסטוגרמיה, דיאגרמת גבעול עלה, דיאגרמת קופסה.

* **מצדים סטטיסטיים:**

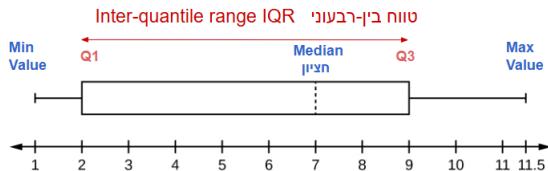
שכיח: הערך עם השכיחות הגבוהה ביותר.

מרכז הטווח: הממוצע בין התצפית הגבוהה ביותר והנמוכה ביותר.

חציוון: 50% או פחות (אם לא זוגי) גובהים ממנה, 50% או פחות נמוכים ממנה.
ממוצע: סכום כל הערכים מחולק במספר התצפיות.

דיאגרמת קופסה:

בשביל לחישבה מסתכלים על ערך המינימום, המקסימום, החציוון וכן Q_1 ושיהיה החציוון של החצי הנמוך (מהמינימום עד החציוון) ו Q_3 ושיהיה החציוון של החצי הגבוה (מהחציוון אל המקסימום).



כיצד מציירים דיאגרמת קופסה?

א. מסדרים את הנתונים לפי סדר עולה

ב. מוצאים מינימום, מקסימום, חציוון, רביעון ראשון ורביעון שלישי

ג. מציירים לפי הנתונים שמצאנו קודם – בין הרביעון הראשון לרביעון השלישי אונחנו מציירים קופסה, בהוכחה מסמנים את החציוון. הטעות שבין הרביעון הראשון לרביעון השני נקרא טווח בין רביעוני שלושה.

ד. לאחר מכן מציירים קוים לקצה הטווח – בין הרביעון הראשון למינימום ובין הרביעון השלישי למקסימום

הערה: **למציאת החציוון – אם יש לנו מס' זוגי זה קל, האמצעי.** אם יש מס' זוגי יש שני חציוונים – החציוון בקורס יוגדר להיות ממוצע שני החציוונים הנ"ל.

3 הרזאה 3: סטטיסטיקה תאורית 2

3.1 מגדדים סטטיטיסטיים

סטטיסטי הסזר: יהיו X_1, \dots, X_n משתנים מקרים עבור אוכלוסייה או מדגם. יהיו x_1, \dots, x_n הערכים שנמדדו עבורם בהתקאה. נסדר את הערכים x_1, \dots, x_n בהתקאה בסדר עולה. קיבל את סטטיסטי הסדר:

$$x_{(1)}, \dots, x_{(n)}$$

שכיח: הערך עם השכיחות הגבוהה ביותר,

$$\bar{x} = argmax\{_{1 \leq i \leq n}(f(x_i))\}$$

מרכז הטווח: הממוצע בין התצפית הנמוכה ביותר לבין הגבוהה ביותר.

$$\bar{x} = \frac{1}{2}(x_{(1)} + x_{(n)})$$

חציוון:

$$\bar{x} = \begin{cases} x_{(k+1)} & n = 2k + 1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & n = 2k \end{cases}$$

ממוצע: סכום הערכים חלקים מס' התצפויות

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

3.2 חישוב מדדי מיקום מרבי עבור מחלקות עם גבולות אמיטיים

כאשר נתונה לנו טבלת שכיחויות וגבולות אמיטיים, נסמן $f(v)$ כשיעור של v , את השכיחות היחסית $rf(v)$ ואת היחסית המצחברת $RF(v)$. נחשב את הממוצע כך:

$$\frac{\sum_v f(v) \times v}{\sum_v f(v)}$$

באשר v הוא מרכז העמודה (אם אנחנו עם גבולות אמיטיים).

כיצד נחשב את החציון בטבלת מחלקות רגילה? נסתכל על השכיחות היחסית המצחברת $RF(v)$ ונחפש היכן אנחנו פחות מחצית מהקלט, והחציון יהיה שורה אחת אחריו. אם קיימים ערך עבורו $RF(v) = 0.5$ אז חציון יהיה הממוצע של זה לפני וזה אחריו.
כיצד נחשב חציון עבור מחלקות עם גבולות אמיטיים?
מחלקה m , גבולות $L_0 - L_1$, שכיחות f ומצטברת F .

$$Md = L_0 + \frac{\frac{n}{2} - F(X_{m-1})}{f(x_m)}(L_1 - L_0)$$

הרעיון יהיה למצוא את האיבר אשר הקו מתחתיו מחלק את ההיסטוגרמה לשני חלקים שווים שטח. בשלב הראשון נctrיך לזרות את המחלקה m בה החציון אמרור להמציא, את זה נעשה כמו שעשווים בטבלה רגילה. נסמן את הגבולות שלה ב- $L_0 - L_1$. וכן n מס' התצפויות.

באופן דומה, לחשב את הרבעונים: נזהה את המחלקה m_1 בה נמצא הרביעון ונחשב $-$ נשים לב כי F הינה שכיחות מצטברת (לא ייחסית!).

$$Q_1 = L_0 + \frac{\frac{n}{4} - F(X_{m_1-1})}{f(x_{m_1})}(L_1 - L_0)$$

$$Q_3 = L_0 + \frac{\frac{3n}{4} - F(X_{m_1-1})}{f(x_{m_1})}(L_1 - L_0)$$

עבור מאון k :

$$C_k = L_0 + \frac{\frac{n \times k}{100} - F(X_{m_1-1})}{f(x_{m_1})} (L_1 - L_0)$$

ואלפין : k

$$C_k = L_0 + \frac{\frac{n \times k}{1000} - F(X_{m_1-1})}{f(x_{m_1})} (L_1 - L_0)$$

הערה. נשים לב כי הנוסחאות הנ"ל תקפות אך ורק כאשר אנחנו מדברים עם גבולות אמתיים (גבול עליון של מחלוקת קודמת זהה לנגבול תחתון של מחלוקת נוכחית).

3.2.1 מהו המדד הכי טוב עבור \bar{x} מיקום מרכז?

אם נבחר במדד מסוים, מהי פונקציית ההפרש של?

א. מס' השגיאות: כמה מוחרכים אינם שווים למדד עצמו $|\{x_i | x_i \neq \bar{x}\}|$:
כאשר נסתכל על פונקציה זו, השכיח ימזרע את מס' השגיאות. כלומר אם הפונקציית הפסד שמעניינת אויה היא מס' השגיאות אויש נשתמש בשכיח.

ב. השגיאה המקסימלית: המרחק המקסימלי מהמדד עצמו $\max_i |x_i - \bar{x}|$
כאשר נסתכל על מדד זה, מרכז הטווח ממזרע את השגיאה המקסימלית.

ג. סכום השגיאות המוחלטות: מרחקים אבסולוטיים של כל הערכים הממדד $\sum_i |x_i - \bar{x}|$
הຊיזון ממזרע את סכום השגיאות המוחלטות.

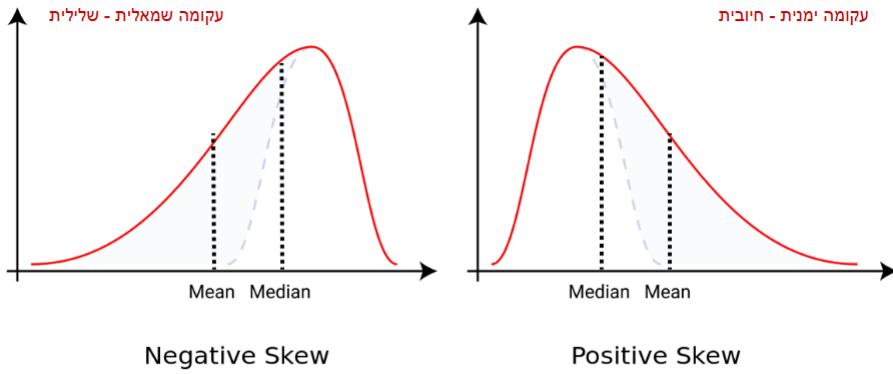
ד. סכום ריבועי השגיאות: מרחקים ריבועיים של כל הערכים מהמדד $\sum_i (x_i - \bar{x})^2$
הממוצע מפחית למינימום את סכום ריבועי השגיאות.

מכאן ניתן כי כל פונקציית הפסד מתיחסת ו"מענישה" מדד אחר. לכל שימוש ישנו מדד שונה שטוב עבור \bar{x} .

תכונות מדדים סטטיסטיים למיקום מרכז \bar{x} :

משמעות	שכיח	אםצע טווח	חציון	ממוצע
פונקציית הפסד	מספר שגיאות	סכום השגיאות המוחלטות	סכום ריבועי השגיאות	סכום ריבועי
رجישות לערכי קיצון	רבה	מעטה	אין	רבה
סולמות מדידה	רוחחים ומעלה	סדר ומעלה	רוחה ומעלה	רוחה ומעלה
שימושות בהסקה	רובה	פחותה	בינונית	רובה

נשים לב. בעקבות פעמון סימטרית: המוצע=חציון=שכיח.
בעקבות פעמון אי סימטרית שמאלית (הציג לצד שמאל) : המוצע $>$ חציון $>$ שכיח
ביקומת פעמון אי סימטרית ימנית (הציג לצד ימין) : המוצע $<$ חציון $<$ שכיח



3.3 מדדי פיזור

- א. **אחוז השגיאות:** אחוז התצפויות השונות מהשਬיך $\frac{1}{n} |\{i | x_i \neq \bar{x}\}|$
- ב. **גודל השגיאה המקסימלית:** המרחק הגדול ביותר ממרכז הטווח $\max_i |x_i - \bar{x}|$
- ג. **הטווח:** המרחק בין ערכי קיצון
- ד. **הטווח רביעוני:** הטווח בו נמצאים 50% הערכים המרכזיים בהתקפלות. (מה שאנוanno מציירים בדיאגרמת Box).
- ה. **ממוצע הסטיות המוחלטות:** ממוצע מרחקי התצפויות מהחץון.
- ו. **ממוצע ריבועי הסטיות:** ממוצע ריבועי מרחקי התצפויות מהממוצע $\frac{1}{n} \sum_{i=1}^n |(x_i - \bar{x})|^2$. נשים לב כי הউലা בربוע מענישה יותר את הקצוות, וזה יותר טוב עבור המר策, ולכן זה בודק את הקצוות.

תכונות:

פונקציית פסד	פונקציית הגחר	לערכי קיצון הчисelog	מהירות הчисלוג	ריגשות לכל הערלים	שכיח	מדד המרכז	מספר שגיאות	אחד השגיאות	שגיאה מקסימלית	טוווח בינרבעוני	ממוצע סטיות ריבועיות	ממוצע סטיות מוחלטות	סכום סטיות ריבועיות	סכום סטיות סטיות	פונקציית ריבועיות
+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
+	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+

3.4 שונות וסטיות תקן

אנו נשתמש בעיקר בממוצע סטיות ריבועיות, הידועה בשם: **שונות**.

עבור רשימת ערכים:

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

וסטיטית התקן:

$$S_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

עבור טבלות שכיחיות:

$$\frac{1}{n} \sum_x (x_i - \bar{x})^2 f(x) = \frac{1}{n} \sum_x x_i^2 f(x) - \bar{x}^2$$

יש לשים לב - השונות וסטיטית התקין באוכולסיה ובמדגם שונים זה מזה. באוכולסיה, כפי שראינו לעיל. במדגים:

$$S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

מדוע מחלקים ב $n-1$ ולא ב n ? נגלה בהרצתה.

שימושים לממוצע וסטיטית התקן: עבור עיקומת פעמו, בערך 68% מהמערכות הם במרחב של סטיטית התקן אחת מהממוצע. בערך 95% מהמערכות הם במרחב של שתי סטיטיות התקן מהממוצע.

חוק צבישב: עבור כל התפלגות,
לפחות 75% מהמערכות הם במרחב 2 סטיטיות התקן מהממוצע.
לפחות 88.89% מהמערכות הם במרחב 3 סטיטיות התקן מהממוצע,
באופן כללי לפחות $1 - \frac{1}{k^2}$ מהמערכות הם במרחב k סטיטיות התקן מהממוצע.

3.5 ממוצע משוקל וסוגנות מצורפת

עבור k כיוות שונות, בהינתן N מס' התלמידים בשכבה מותקיים כי הממוצע המשוקל הינו:

$$\bar{X}_T = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij} = \sum_{j=1}^k \bar{x}_k \times \frac{n_j}{N}$$

עבור k כוונות שונות, השונות המצורפת הינה:

$$S_T^2 = \frac{1}{N} \sum_{j=1}^k \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_T)^2 = \sum_{j=1}^k \frac{n_j}{N} S_j^2 + \sum_{j=1}^k \frac{n_j}{N} (\bar{x}_j - \bar{x}_T)^2$$

החלק הימני בביטוי הוא השונות בין הקבוצות השונות, והחלק השמאלי היא השונות בתחום הקבוצות (סוכמיים).

תיקנו: למשל, סטודנט קיבל 70 בחשבון ו 57 בתנך. היכן הצלחה יותר? בחישוב הממוצע היה 65 וסטיית תקן 3. בתנך 70 ו 41 בהתאם. כיצד נדע? נרמל -

ציוו התקן של x : מרחק מהממוצע הנמדד ביחסות סטיית התקן.

$$z_x = \frac{x - \bar{x}}{S_x}$$

מכאן, נקבל שהתוצאות יהיו בתחום העקומות Z המפורסמת - התפלגות נורמלית סטנדרטית.

3.6 ממדדי קשר בין מספר משתנים

למשל: האם יש קשר בין טמפרטורה ממוצעת באזור לתנועת עצי פרי באזורי?

עד כה דנו במשתנה אחד, נדונו כתם' משתנים.

יהיו לנו n תצפיות ובכל אחד מהתצפיות יש לנו ערכים (x, y) . במערך ניסוי שכמה נרצה ללמידה על הימצאות הקשר בין x ל y .

ונכל להשתמש בדיאגרמת פיזור: על ציר x ערכי x ועל ציר y ערכי y . לבדוק האם קיים קשרلينאר.

מقدم המתאים של פירסום: ממד קשר הממלא אחר הדרישות הבאות -

א. ערכו המוחלט יהיה מקסימלי באשר הקשר מושלם (כל הנקודות על הישר - הקשר לינארי)

ב. סימנו של הממד שלילי או חיובי יבטא את ציוווקן הקשר (חיובי כאשר חיובי ולהפך).

$$r = \frac{\sum_{i=1}^n z_{x_i} z_{y_i}}{n} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}$$

זכור כי הגדרת השונות המשותפת:

$$cov(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

ומכאן ש:

$$r = \frac{cov(x, y)}{s_x s_y}$$

באשר אנחנו עובדים עם מודם אנחנו נחלק ב-1 – a .

נרצה ליצור קו מגמה מהצורה $b = ax + y$. על זאת נלמד – בהרצאה 4.

4 הרצאה 4: הסקה סטטיסטית

4.1 מבוא

נשים לב, מה עשינו עד כה בקורס: שאלנו כיצד חוקרים מגילם? מנחים השערה (הגדרת משתנים וסולמות מדידה), אוסף נתונים (מאוכלסיה ומדגם), מארגנים את הנתונים (הציג טבלאית כמו שכיחות או צפיפות או ויזואלית כמו היסטוגרמָה וכן מחשבים מדדים סטטיסטיים) ולבסוף מסיקים מסקנות. אם הנתונים נאספו על כלל האוכלוסייה אז סיימנו. **אם הנתונים נאספו על מודם מייצג מותו כלל האוכלוסייה: עוד לא סיימנו.**

נרצה לשאול כמה שאלות חשובות. האם ניתן להכליל ממדדים במודם לממדים באוכלוסייה? באיזו רמת בטחון ניתן לבצע הכללה זו? האם לקבב או לדוחות השערה ותחת אילו תנאים? המטרה העיקרית בהרצאה זו תהיה לבדוק – האם יש קשר בין תופעות המודם לתופעות באוכלוסייה? כיצד נעשה זאת: באמצעות הסתברות. נראה כיצד הסתברות וסטטיסטיקה נפשים.

4.2 הסקה סטטיסטית

זכור כי ישנה הגישה המדעית שמורכבת משתי גישות. האמפירית ("הכל מדיד"), והרצינולית: גישה שבבסיסה על כללי היסק.

הסקה דזוקטיבית: ($\text{כל } \leftarrow \text{פרט}$), היסקelogי, אמונות והנחהות מהיבת את אמונות המסקנות. למשל: הנחה¹ אין מים על כוכב הלכת חמה, הנחה² לא מים אין חיים. אז מסקנה: אין חיים על כוכב הלכת חמה. ניתן להפריך את ההנחהות אך לא את המסקנה (!!).

הסקה אינדוקטיבית (פרט $\leftarrow \text{כל}$): הכללה, הנחות מובילות למסקנה בסביבות גבולה. לא מוחלטת. למשל – הנחה: כל הברבורים שנצפו עד היום היו לבנים. מסקנה: הברבור הבא שנראה יהיה לבן. דוגמה נוספת – עד היום השימוש זרחה כל בוקר, אז היא תזרח גם אחר. ניתן להפריך את המסקנה! (!!).

הבעיה המהותית: מה ההצדקה להסקה אינדוקטיבית במדע (כל המדע מותבבס על הסקה שכזו)? לא נלמד זאת בקורס – זה מדעי הדשה. עם זאת, הבעיה המכוטית: כיצד ל证实 את מידת הוודאות שבתוך אי הוודאות? כן בקורס שלנו.

4.3 מושגים בסיסיים

משתנה מקרי: "תמונה" שהיא משתנה שלקורה מהתפלגות מסוימת F . ככלומר $F \sim X$.

תצפית: תוצאה של ניסוי מותוך המשתנה המקרי X .

דוגמה: ביצוע רצף תצפיות (ניסויים) X_1, \dots, X_N אשר $F \sim X_i$ $\forall i$.

מודם מקרי בגודל n מותוך מ"מ X : מודם של n משתנים מקרים כך ש:

א. X_1, \dots, X_n הם מ"מ בלתי תלויים
ב. לכל מ"מ X_i יש את אותה פונקציית ההסתברות כמו של X , ככלומר לכל i מתקיים $F \sim X_i$.

משמעות: דוגימה מקרים (אקרואיד, רנדומית) עם החזירה של n איברים מותוך אוכלוסייה עם תמונה $F \sim X$ שקופה למודם מקרי בגודל n מותוך מ"מ מותאים X . ולהפוך.

מסקנה: מבחינה מעשית (\leftarrow) נבצע דוגימה מקרים מותוך אוכלוסייה גם כאשר בפועל נרצה לדוגם מומ"מ. וכן מבחינה תאורתית (\Rightarrow) נוכל להשתמש בכל מה שאנו ידעים על מ"מ על כל דוגמה מקרים.

ניסי לב: תכונה של האוכלוסיה נקראת פרמטר, וערכו קבוע אך לא בהכרח ידוע. מודד המבוסס על המדגמים נקרא סטטיסטי וערכו ידוע אך לא בהכרח קבוע.
עבור אוכלוסיה בגודל k ניתן לייצר הרבה מדגמים שונים בגודל $k < n$ מכאן שכל סטטיסטי הוא משתנה מקרי עם התפלגות משל עצמו – התפלגות הדגימה של הסטטיסטי. (כלומר, תאסוף את כל המדגמים, כל אחד מהם מוציא סטטיסטי, כל סטטיסטי הוא משתנה מקרי עם התפלגות שלו.).

טענה: הסטטיסטי הוא משתנה מקרי עם התפלגות דגימה: לה נקרא – התפלגות הדגימה של הסטטיסטי.

4.4 התפלגות דגימה

צורת התפלגות הדגימה: צורת ההתפלגות תלויות במספר גורמים – בסוג ההתפלגות באוכלוסיה, בסוג הסטטיסטי ובגודל המדגמים. لكن נקבע כשנדבר על "התפלגות דגימה": התפלגות הדגימה של סטטיסטי מסויים s עבור מדגמים בגודל n שנלקחו מאוכלוסייה בה ערכי המשתנה מתפלגים לפי התפלגות F .

התפלגות הדגימה של הממוצע (ממוצע המדגם): מסומן \bar{X} והוא משתנה מקרי בעל פונקציית הסתברות ונitin לחשב עבורי תוחלת ושותות.

משפט: תוחלת הסטטיסטי "ממוצע המדגם" (ממוצע כל המדגמים) \bar{x} שווה לתוחלת המ"מ X ממנו אנו דוגמים. ככלומר $E[\bar{x}] = \mu_x$.
הוכחה:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$E[\bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = E[x_i] = E[X]$$

טענה: $(\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}) \quad Var[\bar{X}] = \frac{Var[X]}{n}$
הוכחה:

$$Var[\bar{X}] = Var\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n^2} \sum_{i=1}^n Var[x_i] = \frac{1}{n^2} \times n \times Var[X] = \frac{Var[X]}{n}$$

$$\text{מסקנה: } \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$$

תזכורת: אם $X \sim N(\mu, \sigma^2)$ כך שגם היא התוחלת μ והיא סטיית התקן. (σ^2 היא השונות).

מודע זה מוכח למשתמש מדגמים אחד? ברור כי במדגים שונים עבור אותה אוכלוסייה יש ממוצעים שונים, אם נדemos הרבה מדגים ונחשב ממוצע לכל אחד, ממוצע הממוצעים יתרקרב מאוד לממוצע באוכלוסייה. **אבל:** אין בכוונתו לדגום הרבה מדגים! אלא מגם אחד ויחיד! אז: השאלה למעשה: מהי הסבירות שהממוצע במדגם שדגמנו סוטה (בהרבה) מהתמוצע באוכלוסייה? שאלת שוקלה: מהי

הסבירות שהמוצע במדגם שדגמנו סוטה בהרבה מהתוחלת של הממם (ממוצע המדגם) עצמו? ככלומר - כמה הערך שלי רחוק מהתוחלת של ממוצע המדגם. זו שאלה עדיפה לנו - כי אכן יש מדגם אחד בדיק. לפיכך: נתענין במידת הפior של התפלגות הדגימה של ממוצע המדגם. סטיית התקן של ממוצע המדגם שווה לסטיית התקן של הממם המקורי מוחלקת בשורש n גודל מדגם ולכן: **ככל שהמדגם גדול יותר, שונות/סטיית התקן של ממוצע המדגם תהיה קטנה יותר**

מסקנה - נרצה שהשונות וסטיית התקן תהיה קטנה מאוד ולכן ככל שהמדגם גדול יותר כך השונות וסטיית התקן יהיו קטנות. לכן - נרצה מדגם יחיד גדול.

הוכחה:
לפי אי שוויון צביש'ב מתקיים

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

אם נפעילו על הממוצע \bar{X} נקבל

$$P\left(\mu - k \frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + k \frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}$$

באשר n שואף לאנסוף, נראה כי ממוצע המדגם כלוא בין שני ערכי μ ולכן שווה לו.
נבחר $\varepsilon \leq \frac{\sqrt{n}}{\sigma}$ ונציגו, נקבל

$$P(\mu - \varepsilon < \bar{X} < \mu + \varepsilon) \geq 1 - \frac{\sigma^2}{\varepsilon^2 n}$$

ומכאן נקבל את **חוק המספרים הגדולים**:

$$\lim_{n \rightarrow \infty} P(\mu - \varepsilon < \bar{X} < \mu + \varepsilon) = 1$$

כלומר: אם נkeh הרבה מאוד תצפיות, כאשר מס' התצפיות שואף לאנסוף נקבל כי ההסתברות שסכום הממוצעים שווה לתוחלת היא 1.

טענה: בדגימות מודגם שגודלו n מתיוך ממ X המתפלג נורמלית עם תוחלת μ וסטיית התקן σ יהיה ממוצע המדגם \bar{X} גם הוא ממ המתפלג נורמלית עם תוחלת μ וסטיית התקן $\frac{\sigma}{\sqrt{n}}$.

משפט הגבול המרכזי: נסמן $S_n = \bar{X}$. נתונים X_1, \dots, X_n משתנים בלתי תלויים זהים (כלומר עם אותה התפלגות) עם תוחלת μ ושונות σ^2 . ככלומר לכל $n \leq i \leq 1$ מתקיים $E[X_i] = \mu$. כך $.X_1 + X_2 + \dots + X_n = S_n$ שנדיב

$$Z_n = \frac{S_n - \mu n}{\sigma \sqrt{n}}$$

$$\text{מתקיים } .Var[Z_n] = 1 \text{ ו } E[Z_n] = 0$$

אז, יהא $Z \sim N(0, 1)$

$$\forall z : \lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z) = \Phi(z)$$

הערה חשובה: מודגם "גדול משפיק" הוא זה בגודל שגודלו $30 \geq n$. ורק אם הוא בגודל שגודלו מ-30 אפשר להשתמש במשפט הגבול המרツצי.

5 הרצאה 5: אמידה סטטיסטית נקודתית

היכן אנחנו כתע נמצאים? לומדים הרסקה סטטיסטית. נושא זה מתחלק ל-2: אמידת פרמטרים (הרצאה 6 – 5) ובדיקת השערות (הרצאה 9 – 7). אמידת פרמטרים מוחלקת ל-2: בהרצאה זו נדבר על אמידה סטטיסטית נקודתית ובהרצאה הבאה נדבר על אמידת מרוחקי בטחון. היום נדבר על השאלה הבאה: כיצד והאם ניתן להכליל נתונים למצאים באוכלוסייה?

פרמטר: גודל קבוע המופיע את כל האוכלוסייה.

סטטיסטי: ערך המוחושב ע"פ המודגמים.

אמידה היא הערכת (שעוריך) ערך הפרמטר ע"פ סטטיסטי המודגמים.

5.1 אמידה סטטיסטית

ישנן שתי שיטות לערכית אמידה סטטיסטית.

1. **אמידה נקודתית:** ההנסחה המומוצעת של משפחה בת 4 נפשות היא 11,500 שקלים בחודש – על סמך המדגם מחושב סטטיסטי אחד.
2. **רווח סמי:** בהסתברות של 80% ההנסחה המומוצעת של משפחה בת 4 נפשות בישראל היא בין 8000 ש"ח ל-16,000 ש"ח – על סמך המדגם מחושב טווח של ערבים.

לאמידה סטטיסטית נקודתית ישן בעיות:

- א. בעיה מהותית – הסטטיסטי הוא רק אומדן. כיצד נדע את ערך הפרמטר ביחס לאוכלוסייה כולה? (לא נדע). מדוע מותר להשתמש בהנסחה אינדוקטיבית? (לא בקורס זה).
- ב. בעיה מעשית – בעיה כמותית, באיזה סטטיסטי כדי ליהשתמש כדי לאמוד משתנה מסוים? איזה אמדים קיימים ואיזה תכונות יש להם? מה נדרש לאמוד טוב?

5.2 בעיית האמידה

נתון: עברו משתנה מקרי $F \sim X$ ונתון מוגם מקרי x_n, x_{n-1}, \dots, x_1 ב"ת באשר $F \sim \sum_{i=1}^n x_i$.

הנחת בעודה: אנו ידועים את צורת ההתפלגות של X – פונקציית ההסתברות או הצפיפות אך לא ידועים את הפרמטר.

בעיית האמידה: מהם ערכי הפרמטרים של פונקציית ההסתברות או הצפיפות $\sim F$.

דוגמא: אמידת זמן חיים של נורה (λ) $\sim \exp(\lambda)$. אמידת פרמטרים של ההתפלגות נורמלית גבוהה או משקל של בניים או בניית (μ, σ^2) .

טרמינולוגיה:

עבור פרמטר באוכלוסייה θ נסמן את האמד במדגם $\hat{\theta}$.

דוגמא: עבור התוחלת μ נבחר אמד שיחיה המומוצע: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$. (מטרת השיעור היא להסביר מדוע בהכרח הממוצע היא האמד הכי טוב לתוחלת. זה מוכח שזה האמד הכי טוב שיש. בהמשך נראה הוכחה לכך). הדגשה חשובה – אין לי מושג מה ערכה של μ באוכלוסייה. אך יש לי מדגם. אני רוצה להסיק על התוחלת, באמצעות המדגם ולכן מחשבים את האמד $\hat{\mu}$

אבחןות חשובות: לאותו אמד נקבל תוצאות שונות על מוגדים שונים. מכאן שהאמד (הסטטיסטי) הוא בעצמו משתנה מקרי. ומכאן שלאמד (הסטטיסטי) עצמו יש התפלגות דגימה. מה שיכתיב את התכונות של האמד תהיה התפלגות הדגימה של הסטטיסטי.

- הגדרה:** נתון מוגד מקרי X_n, X_1, \dots, X_n . אנו רוצים לאמוד את ערכי θ מתוך המוגדים. אזי,
1. פרמטר - פונקציה של ערכי האוכלוסייה. יכולה להיות תליה בפרמטרים לא ידועים.
 2. סטטיסטי - פונקציה של ערכי המוגדים. אינה תליה בפרמטרים לא ידועים.
- דוגמה:** $\sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n x_i$ (\bar{X}) הוא סטטיסטי. אך $\sum_{i=1}^n (x_i - \mu)^2$ אינו סטטיסטי כי תליה בעמ.
3. אמד (*estimator*) - סטטיסטי שבאזורו אומדים פרמטר בלתי ידוע (פונקציה כללית).
- לדוגמה: המוצע הוא אמד לתוחלת. **בשאנו מעריכים אמד:** אנו מעריכים נסחה.
4. אומדן - המספר עצמו שמציבים בנסחה (אמד) עבור מקרה ספציפי. התוצאה שקיבלנו עבור האמד במוגדים ספציפי (תוצאה ספציפית).
- דוגמה: המוצע במוגדים הטלה קוביה 1, 2, 1, 4, 2, 6, 4, 2, 5 הוא האומדן במוגדים:

$$\hat{\mu} = \frac{1 + \dots + 5}{9} = 3$$

5. שגיאת האמידה - המרחק בין ערך האמד לערך הפרמטר: $\theta - \hat{\theta}$. נשים לב כי את θ איןנו ידעים. אז כיצד יעזר לנו לחשב ערך אמד (במוגדים ספציפי)? אנחנו נרצה לחסום ככל שניתן את שגיאת האמידה.
6. הטיה של אמד - התוחלת של שגיאת האמידה

$$E[\hat{\theta} - \theta] = E[\hat{\theta}] - E[\theta] = E[\hat{\theta}] - \theta$$

שכן הערך של θ הינו קבוע ושל $\hat{\theta}$ אינו קבוע. מכאן נגדיר רשותה שההטיה של אמד הינה:

$$Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta$$

5.3 תכונות של אמידים

מהו אמד שהוא טוב?

1. **אמד עקי** - ככל שהמוגדים גדול ההסתברות שהאמד יתכנס לפרמטר האמייתי גדול. כלומר: $\hat{\theta} \rightarrow_{n \rightarrow \infty} \theta$
2. **חסר הטיה** - הנטיה של האמד שווה לאפס. כלומר, $E[\hat{\theta}] = \theta \iff Bias(\hat{\theta}, \theta) = 0$. כלומר: אם אמדנו הרבה פעמים, והיו לנו שגיאות מהמדד האמייתי בכל אחת מהדגימות אך בתוחלת השגיאות הללו ביטלו אחת את השניה והתקרבנו לממד האמייתי.

תכונות ממוצע המוגדים:

- עבור תוחלת μ נגדיר את ממוצע המוגדים כאמד: $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$
- האם המוצע של המוגדים הוא אמד טוב לתוחלת?
- א. אכן אמד עקי - ככל שהמוגדים גדול, ערך האמד \bar{X} מתכנס לערך הפרמטר באוכלוסייה. זה מגע בבדיקה המספריים שלמים:

$$\lim_{n \rightarrow \infty} P(|\bar{X} - \mu| < \varepsilon) = 1$$

ב. אכן חסר הטיה - ראיינו כי $E[\bar{X}] = E[X_i]$ בהערכתה הקודמת (באשר זה תוחלת של מוגדים כלשהו), ומכאן שנקבל כי אכן $Bias(\hat{\theta}, \theta) = 0$

טענה (עבור כל התפלגות): בהינתן מוגדים מקרי x_1, \dots, x_n ב"ת מתוך מ"מ X עם תוחלת μ ומשונות σ^2

א. אמד לתחוללת שהוא חסר הטיה הוא הממוצע $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n x_i}{n}$

ב. אמד לשונות (בהינתן שהתחוללת ידועה!!!!) שהוא חסר הטיה: $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$

הוכחה: של א':

$$\hat{\mu} = E[\bar{x}] = E\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E[x_i] = E[x_i] = E[X] = \mu$$

של ב':

$$\hat{\sigma}^2 = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2\right] = \frac{1}{n} E\left[\sum_{i=1}^n (X_i - \mu)^2\right] = \frac{1}{n} \sum_{i=1}^n E[(X_i - \mu)^2] = \frac{1}{n} \sum_{i=1}^n \sigma^2 = \frac{1}{n} \times n \times \sigma^2 = \sigma^2$$

בעתណזון באמד לשונות עם תוחלת שאינה ידועה. אם אין לנו תוחלת, אולי כדאי להסתכל על הממוצע \bar{X} ?

$$\hat{\sigma}^2 = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n} E\left[\sum_{i=1}^n (X_i - \bar{X})^2\right] = \frac{1}{n} \sum_{i=1}^n E[(X_i - \bar{X})^2]$$

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2 =$$

$$\sum_{i=1}^n (X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 =$$

$$\sum_{i=1}^n (X_i - \mu)^2 - 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \mu) + \sum_{i=1}^n (\bar{X} - \mu)^2 =$$

$$\sum_{i=1}^n (X_i - \mu)^2 - (*) 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2 = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2$$

באשר (*) מחייב הסביר: מדובר על $n(\bar{X} - \mu)$
כעת:

$$E[(X_i - \bar{X})^2] = E\left[\sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2\right] =$$

$$\sum_{i=1}^n E[(X_i - \mu)^2] - nE[(\bar{X} - \mu)^2] = \sum_{i=1}^n \sigma^2 - nVar[\bar{X}]$$

$$= n\sigma^2 - n\frac{\sigma^2}{n} = n\sigma^2 - \sigma^2 = \sigma^2(n-1)$$

ולכן,

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

מסקנה: הממוצע \bar{X} (באשר התוחלת אינה ידועה) הוא מוטה עבור השונות.

לשם כך אנו משתמשים בתיקון בסל: אנו מכפילים את האמד ב- $\frac{n}{n-1}$ ומתקבלים $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

טענה: אמד חסר הטיה לשונות באשר התוחלת אינה ידועה הינו:
א. עבור אוכלוסייה (כי יודעים את התוחלת בהכרח):

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

ב. עבור מבחן (לא יודעים את התוחלת, ומשתמשים בתיקון בסל):

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

טענה: אם θ_1 ו- θ_2 הם אמדים חסרי הטיה עבור θ אז נעדיף את זה עם השונות הקטנה יותר.
את θ_2 המקיים $V(\theta_2) < V(\theta_1)$

יעילות של אמידה: במקרה הכללי - תוחלת ריבועי השגיאות הינה

$$MSE(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$$

אם θ_2 ו- θ_1 הם אמידים שאינם חסרי הטיה עבור θ נעדיף את האומד θ_2 המקיים $MSE(\theta_2) < MSE(\theta_1)$

$$\text{טענה: } MSE(\hat{\theta}) = Var_{\theta}(\hat{\theta}) + Bias_{\theta}(\hat{\theta}, \theta)^2$$

5.4 שיטות אמידה

5.4.1 שיטת המומנטים

שיטת המומנטים היא שיטה אמידה על פי פרמטרים המאפיינים התפלגות של אוכלוסייה מסוימת. נניח מושתנה מקרי המתפלג F עבורו ישנו k פרמטרים בלתי ידועים נגיד **פונקציה מייצרת מומנטים** (mfg). נאמדות את המומנט ה- k באמצעות מוצע חזקה ה- k של התציפות.

$$\mu_1 = E[X], \mu_2 = E[X^2], \mu_3 = E[X^3], \dots, \mu_k = E[X^k]$$

נראה כי μ_1 הוא מרכז הנתונים, μ_2 הוא דומה ומזכיר את השונות (הפייזר), μ_3 מעיד על לאיזה כיוון העקומה הולכת, μ_4 מעיד על עובי האנבות" וכך זה ממשיך.

כל אמד למומנט מחושב כך לפי ערכי x_1, \dots, x_k שהושבו במדגם.

$$\mu_k = \frac{\sum_{i=1}^n x_i^k}{n}$$

השיטה אומرت כך:

א. נשווה כל מומנט מסדר k לאומדן שלו במדגם:

$$\mu_1 = g_1(\hat{\theta}_1, \dots, \hat{\theta}_n)$$

$$\mu_2 = g_2(\hat{\theta}_1, \dots, \hat{\theta}_n)$$

...

$$\mu_k = g_k(\hat{\theta}_1, \dots, \hat{\theta}_n)$$

ב. פותרים את מערכת המשוואות של k המשוואות ב- k הנעלמים.

דוגמה:

נניח כי X מותפלג מעריכית ($X \sim Exp(\lambda)$). המומנט הראשון הינו התוחלת $E[X] = \frac{1}{\lambda}$. האמד של המומנט הראשון הינו $\frac{\sum_{i=1}^n x_i}{n}$ (הממווצע). מכאן משווים: $\frac{\sum_{i=1}^n x_i}{n} = \hat{\lambda}$ ומקבלים $\hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n}$ (למה שמנו כובע? זה אמד. אנחנו לא יודעים מה ערכו בבדיקה של λ).

הספיק לנו מומנט ראשון כי רצינו למצוא משתנה יחיד. נתבונן בדוגמה נוספת:
נניח משתנה מותפלג אחיד $X \sim U[a, b]$. אז משווהה ראשונה לפי המומנט הראשון:

$$\frac{\sum_{i=1}^n x_i}{n} = \frac{\hat{a} + \hat{b}}{2}$$

משווהה שנייה לפי המומנט השני:

$$\frac{\sum_{i=1}^n x_i^2}{n} = E[X^2] = Var[X] + (E[X])^2 = \frac{(\hat{b} - \hat{a})^2}{12} + \frac{(\hat{a} + \hat{b})^2}{4}$$

סה"כ קיبلנו שתי משוואות שני נעלמים. הרি x נתונים לנו וגם a . נקבל

$$\hat{a} = \bar{X} - 3 \frac{\sum_{i=1}^n x_i^2}{n} - 3\bar{X}^2$$

$$\hat{b} = \bar{X} + 3 \frac{\sum_{i=1}^n x_i^2}{n} + 3\bar{X}^2$$

וכך, מנתונים שידיוע שמתפלגים בצורה איחידה, הצלחנו למצוא אמד ל- a ול- b הנדרשים.

יתרונות השיטה: קל להרישוב, נוחה, ניתנה לחישוב עבור כל צורת התפלגות.
 חסרונות השיטה: אם יש הרבה פרמטרים זה יהיה לא קל לחישוב, עלולים לקבל אמד מוטה, או אמד שלא נראה סביר.

5.4.2 שיטת הנראות המרבית

נניח שהטלתי מטבע מס' פעמיים. בכל הרטלות קיבלתי 5 (זה ניסוי ברנולי). לפי מה שזה נראה - נראה כאילו $p(X = 5) = 1$.

פונקציית הנראות L : בהינתן ערך p ניתן לחשב את פונקציית הנראות.

נראה כי בהינתן משתנים בינהם, נניח ואנו ידעים כי ההסתברות שיצא 7 פעמיים אותו מספר היא 0.12. כלומר $P(k|n, p) = \dots$. נרצה להפוך אותה לפונקציית נראות $L(p|k, n)$.

כיצד נדע איזה ערך ימקסם את L ? נגזר אותה לפי p ונשווה לאפס.

$$L(p|k, n) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\ln L(p|k, n) = \ln(\binom{n}{k}) + k \ln(p) + (n-k) \ln(1-p)$$

$$\ln L(p|k, n)' = \frac{k}{p} - \frac{n-k}{1-p} \implies \hat{p} = \frac{k}{n}$$

נשים לב. מודיעו המרנו *לעת?* הרבה יותר קל לגזר כך.
שנייה: קיבלנו הוכחה מעניינת לכך **שהשכיחות היחסית היא אמד נראות מרבית עבור פרמטר p בהתפלגות בינומית.**

להלן השיטה:
א. נגידר את פונקציית הנראות של θ כמכפלת ההסתברויות $x_1, \dots, x_n \sim F$ בהינתן θ :

$$L(\theta, X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P(X_1 = x_1|\theta) \times \dots \times P(X_n = x_n|\theta) = \prod_{i=1}^n P(X_i = x_i|\theta)$$

ב. נגידר את לוג פונקציית הנראות

$$\ln L(\theta, X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = \dots = \sum_{i=1}^n \ln(P(X_i = x_i|\theta))$$

ג. נגזר את $\ln L$ לפי θ ונשווה לאפס למציאת ערך קיצון.
ד. נגזר את $\ln L$ פעמייה לודא מקרים.

מינוח. בשאלת של "מצא אן"ם" עושים את שיטה זו.

מודיע אן"ם טוב לנו?

א. **עקביות:** ככל שהمدגם גדול, ערך האמד מתקרב לערך הפרמטר.
ב. **איינוריאנטיות פונקציונלית:** אם θ אן"ם ו- $g(\theta)$ איזי גם $g(\theta)$ אן"ם
ג. **נשים לב - לא ידוע האם האן"ם הוא חסר הטיה**

לשנו:

מודל תאורטי	התפלגות מ"מ	אמד נראות מירבית – אג"ם	האם חסר-הטיה – אג"ה
בינומי	$X \sim \text{Bin}(p)$	$= X/h$	כן
אחדיתה	$X \sim U(1,b)$	$= \max\{X_1 \dots X_n\}$	לא
פואסוני	$X \sim P(\lambda)$	$= \bar{X}$	כן
גאומטרי	$X \sim G(p)$	$= 1/\bar{X}$	לא
מעריצי	$X \sim \text{Exp}(\theta)$	$\theta = 1/\bar{X}$ $\mu = 1/\bar{X}$	לא כן
נורמלית: תוחלת - שונות -	$X \sim N(\mu, \sigma^2)$	$\bar{X} = \frac{\sum_i (X_i - \bar{X})^2}{n}$	כן לא

6 הרצאה 6: אמידה סטטיסטית של מרוחוי בטחון

היכן אנחנו? לומדים תהילך ניסוי, אנו בחלק של אמידת פרמטרים + בדיקת השערות, נשאלו קראנו סטטיסטיקה היסקית.

ראינו כי סטטיסטיקה היסקית מתחלת ב-2:

- א. אמידת פרמטרים: אמידה נקודתית (שיטת המומנטים ושיטות הנראות המרבית) ומרוחוי בטחון - **נושא ההרצאה הנוכחית**.
- ב. בדיקת השערות - בהרמשך.

6.1 רוח סמך של ממוצע המדגמים

נתבונן בממוצע המדגם \bar{X} כאמד נקודתי לתוחלת μ . שגיאת האמידה של ממוצע המדגם היא $\mu - \bar{X}$.
נשים לב כי שגיאת האמידה לא ידועה לנו, ושגיאת האמידה היא משתנה מקרי בעצמה. תחת תנאים אלו נרצה לבדוק את דיקט האמד. דיקט האמד לא יכול להיות מודד אבסולוטי - אלא מודד הסתברותי.
לכן נשאל: מהי הסתברות לכך ששגיאת האמידה של האמד תהיה גדולה מ(\bar{X} בטחון)?
זכור כי לכל $n \geq 30$ עבר משתחה מקרי X בעל תוחלת μ ושונות σ^2 מתקיים $\bar{X} \sim N(\mu, \sigma^2)$ (משפט הגבול המרכזי).

דוגמה.

במדגמים שוגדלו 25 מותוק מ"מ X המתפלג נורמלית בעל סטטיסטיקת תקן $Z = \frac{\bar{X} - \mu}{\sigma}$ לתוחלת μ לא ידועה, מהי הסתברות שסכום המדגמים יהיה שונה מהתוחלת ללא יותר מ-4 יחידות?
נראה כי נתון ($\bar{X} \sim N(\mu, 4)$ ו- $n = 25$, לכן $Z = \frac{\bar{X} - \mu}{\sqrt{4}} \sim N(0, 1)$) שונה ב-4 יחידות = השונות היא $(4\sigma)^2$
כלומר, נדרש

$$P(|\bar{X} - \mu| < 4) = P(\mu - 4 < \bar{X} < \mu + 4) = \int_{Z=\frac{\bar{X}-\mu}{\sigma}} P\left(\frac{\mu - 4 - \mu}{\sigma} < Z < \frac{\mu + 4 - \mu}{\sigma}\right) =$$

$$P(-2 < Z < 2) = \phi(2) - \phi(-2) = \phi(2) - (1 - \phi(2)) = 2\phi(2) - 1 = 0.95$$

מה המשמעות של נתון שכזה? אם נזכיר בגרף ההתפלגות הנורמלית, המשמעות היא שהשיטה שמתחetta לפונקציית הצפיפות הוא 95% והזנות כל אחת 2.5%. כמובן - אם נבצע מוגדים רבים (אנסוי) אז ב-95% מהמקרים ממוצע המוגדים ייפול בתחום זה שנדרש. נשים לב כי את אי השוויון ממנו התחלנו ניתן להמיר לאי שוויון הבא:

$$P(\bar{X} - 4 < \mu < \bar{X} + 4) = 0.95$$

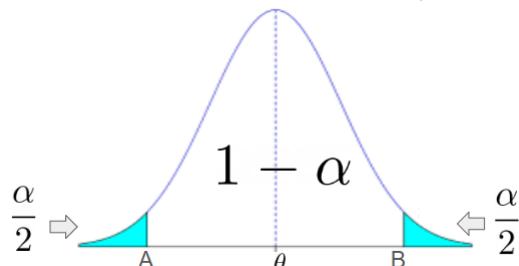
ישנה שיקילות בগל הערך המוחלט. המשמעות לשינוי זה היא - במדגם שגודלו $n = 25$ עברו מ"מ המתפלג נורמלית עם סטיית תקן 4 ותווחת μ לא ידועה, בהסתברות 0.95 הרוחה יכול את μ . כמובן - מצאנו אינפומציה חשובה באשר ל μ . נקרה לביטוי זה: **רווח הסמך של μ ברמה של 0.95.**

6.2 רוח סמך - הגדרה פורמלית

הרוחה (A, B) הוא רוח סמך ברמה של $1 - \alpha$ עבור θ :

$$P(A < \theta < B) = 1 - \alpha$$

כך זה נראה בגרף:



התוצאות האפשריות - היא α (הטורקי) ורמת הסמך היא החלק הלבן.

6.3 רוח סמך לממוצע עבור התפלגות נורמלית כאשר השונות ידועה כאמור לתוחלת

עבור X בעל תוחלת μ , שונות σ^2 וגודל מדגם $n \geq 30$, וכן עבור X נורמלי תוחלת μ , שונות σ^2 וגודל $n > 0$ מתקיים:

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

$$P(-z < X_Z < z) = \phi(z) - \phi(-z) = 2\phi(z) - 1$$

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

נראה כי בшибיל שאנו ימין (הסתברות) $P(-z < X_Z < z)$ תהיה שווה לרוח הסמך ברמה של $:1 - \alpha$

$$2\phi(z) - 1 = 1 - \alpha \implies \phi(Z) = 1 - \frac{\alpha}{2} = z_{1-\frac{\alpha}{2}} = z_{\frac{\alpha}{2}}$$

מעתה נסמן זאת בשם $z_{\frac{\alpha}{2}}$.

דוגמא. אם $\alpha = 0.05$ אז $\alpha = 0.025$ ונקבל $\frac{\alpha}{2} = 0.025 = 0.9750 = 1 - 0.025 = 0.9750$, נלך לחפש ערך זה (0.9750) בטבלת התפלגות הנורמלית Z . נראה כי 1.96 מנייב הסתברות זו כלומר $Z = 1.96$ ולכן $Z = 1.96$ רוח סמך של 95% הוא כאשר $\alpha = 0.05$ כלומר $z_{\frac{\alpha}{2}} = z_{0.05} = 1 - 0.05 = 0.95$ ולכן $\alpha = 0.1$ וערך זה $Z = 1.645$ מתקבל עבור

וכעת, רוח סמך ברמת בטחון של $1 - \alpha$ באשר השונות ידועה הינו:

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

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

דוגמא 2. אורך החכים של נורות מתוצרת מפעל הניצוץ מתפלג נורמלית עם תוחלת μ וסטיית תקן 2.2 נדגמו רנדומית 16 נורות ונמצא שאורך החכים הממוצע הוא 863 . מצא רוח סמך לערך μ של 90% .
פתרון: נראה כי מתקיים $\alpha = 0.1$ ולכן $z_{0.95} = 1.645 = z_{0.95} = 1.645$ ואם נציג בנוסחת רוח הסמך את הנתונים:

$$P(863 - 1.645 \times \frac{22}{\sqrt{16}} < \mu < 863 + 1.645 \times \frac{22}{\sqrt{16}}) = 0.9$$

$$P(853.9525 < \mu < 872.0475) = 0.9$$

6.4 רוח סמך לממוצע כאמד לתוחלת באשר השונות אינה ידועה

נשים לב כי ברוב המקרים השונות לא ידועה לנו מראש. לכן נctrיך לאמוד את השונות σ^2 בעזרת אמד חסר הטיה, כפי שראינו בהרצאה 5.

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

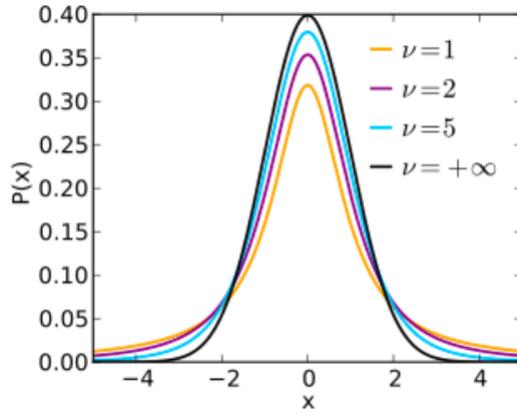
כעת, שינויו את התפלגות הדגימה. כיצד יתפלג המ"מ החדש? עבור $n \geq 30$ מוגדים

לכן, רוח הסמך ברמת סמך (רמת בטחון) של $\alpha - 1$ אחות עברו μ כאשר השונות אינה ידועה עבור מדגמים $n \geq 30$ היא:

$$\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$$

ומה באשר למדגמים קטנים? עבור מדגמים קטנים $n < 30$ נסמננו $t(v)$

התפלגות t :
מהו $t(v)$ שראינו בנוסחה? ישנה התפלגות t שנראית כך:



דרגת חופש הינה כמה מספרים יכולים להשתנות באופן חופשי בהינתן הגבלה מסוימת. למשל בהינתן 5 מספרים וממוצע, 4 יכולים להשתנות מה שהם ירצו להיות אך האחרון חייב להתאים את הערך הכלול בממוצע. לכן דרגת החופש של 5 המספרים הוא 4.

התפלגות זו מתקרבת ל- Z (נורמלית) ככל שיש יותר דרגות חופש.
כל אמד מורייד לנו דרגת חופש אחת, כיון שתליים עוד ועוד. אנו נסמכים על חישוב הממוצע, ולכן דרגת החופש v היא גודל המדגםՓחות אחד. ככלומר,

עבור מדגמים קטנים $n < 30$ נסמננו $T_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim t(n-1)$ רוח הסמך ברמת בטחון של $1 - \alpha$:

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

דוגמה. במדגם בגודל $n = 9$ מאוכטוסיה מתפלגת נורמלית נמצא הממוצע 114. האומדן לסתית 95% התקן באוכטוסיה הינו 12. מצא רוח סמך לתוחלת ברמת סמך של $.95\%$
פתרונות: נתון לנו $n = 9 < 30$ לכן נשתמש בטבלת התפלגות t . וכן $v = 9 - 1 = 8$ ו- $t_{0.025}(8) = 2.306$ מכיוון $\alpha = 0.05$ וכן $t_{0.025}(8) = 2.306(8) = 2.306$

$$114 - 2.306 \frac{12}{\sqrt{9}} < \mu < 114 + 2.306 \frac{12}{\sqrt{9}}$$

סיכום:

נסכם את שאמרנו על רוח סמך עבור ממוצע המדגם עד כה כאשר \bar{X} מתפלג נורמלית.

- ממוצע המדגם הוא אמד עקבי וחסר הטיה לתוחלת

- בחישוב **רוח סמך** עבור התוחלת ברמת סמך α ובשונות ידועה:

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

- בחישוב **רוח סמך** עבור התוחלת ברמת סמך α בשונות **לא** ידועה מדגמים **גדולים**:

$$\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$$

- בחישוב **רוח סמך** עבור התוחלת ברמת סמך α בשונות **לא** ידועה מדגמים **קטנים**:

$$\bar{X} - t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{S}}{\sqrt{n}} < \mu < \bar{X} + t_{1-\frac{\alpha}{2}, n-1} \frac{\hat{S}}{\sqrt{n}}$$

נשים לב, מרוווח הטעות הינו $ME = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ ומתקיים

7 הרצאה 8 + 7: הסקת סטטיסטית - בדיקת השערות

בהרצתה הקרובה נדונן על בדיקת השערות במשתנה אחת וקבוצות קטגוריאליות: הרעיון הכללי וריאציות שונות על אותו נושא - משתנים מסוימים שונים, השוואות שונות.

7.1 השערת האפס וההשערה האלטרנטטיבית

השערה: השערה היא הנחה, היא רעיון שמצוע לשם טיעון כך שניתן יהיה לבדוק אותו כדי לראות אם הוא עשוי להיות נכון.

נסמן את ההשערה הראשונית - **השערת האפס ב** H_0 , **ואת ההשערה האלטרנטטיבית ב** H_1 .
לרוב H_0 היא שאין הבדל בין המדגמים ו H_1 היא שיש הבדל מסוימים (בממוצע, בתפלגות, וכו').

לדוגמא: אם נרצה לדון בשאלת "האם גובהו הממוצע של גברים אסיאתיים גדול משל שאר הגברים באוכלוסייה"? אז אם נסמן μ_{Asian} כממוצע הגובה של גברים אסיאתיים ו $\mu_{Non-Asian}$ כממוצע הגובה של גברים שאינם אסיאתיים אז:

$$H_0 = \mu_{Asian} = \mu_{Non-Asian}$$

$$H_1 = \mu_{Asian} \neq \mu_{Non-Asian}$$

נשים לב - אנו שואלים על הממוצעים. יכולנו לשאול על דברים אחרים כמו התפלגות, סטיטית תקן או חציון.

איך נחליט איזו מההיפותזות נכונה?

קיימות שתי גישות בסיסיות לקבלה החלטה:

א. חישוב הסבירות של השערת האפס H_0 .

ב. גישת קבלת החלטות (מאזור השגיאה).

נשים לב: ההחלטה יכולה להיות מוטעית, אך נשתדר כי הסיכוי לכך יהיה קטן ככל הנימן. וכן: אנו רק מנסים לשלול את השערת האפס - לא להוכיח שההשערה האלטרנטטיבית נכונה.

7.2 סוגים של השגיאות

הגדירה: דחיתת H_0 היא מצב בו הנתונים שנאספו במדגם מספקים ראיות מספיקות (ברמת מובהקות שנקבעה) כדי להסיק שהשערת האפס כנראה אינה נכונה באוכטולסיה.

הגדרה: קבלת H_0 משמעותה היא שהנתונים שנאספו **איןם** מספקים ראיות מספיקות כדי לדוחות את השערת האפס. זה לא אומר בהכרח ש- H_0 נכונה, אלא שאין לנו מספיק הוכחות במדגם כדי לטעון שהיא נכונה.

שגיאה מסוג 1 : דחיתת H_0 בטעות - יש לה שמות נוספים כגון α , α – *False positive*.
קוראים *true negative*

שגיאה מסוג 2 : קבלת H_0 בטעות - יש לה שמות נוספים כגון β , β – *False negative*.
קוראים *true positive*

(True Positive) החלטה נכוןת H_0	שגיאה מסוג I החלטה נכונה	החלטה נכונה בנסיבות H_0 (True Negative)	שגיאה מסוג II החלטה נכונה	ההחלטה על סמך המדגם
דוחית H_0	אי-דוחית H_0			

7.3 חישוב הסבירות של השערת האפס

נניח שאנו רוצים לבדוק אם ממוצע המדגם של גובהים (\bar{X}) שונה מהממוצע של האוכלוסייה (μ) בגבהים. נרצה שמדד עבור הסבירות של השערת האפס יהיה קטן יותר ככל שההבדל בין μ_0 לבין μ יהיה יותר (שהסבירות שההפס נכונה – יהיה לא גבוהה אם הם רוחקים). אפשר לחשב מدد כזה ע"י פונקציה של סטיית התקן סביר המופיע:

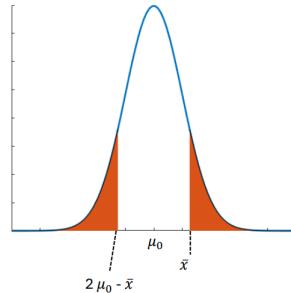
$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

במקרה שהמשתנים נורמליים, נאמר כי הסבירות להשערת האפס תחושב כך:

$$P(\mu_0) = 2(1 - \phi(\frac{|\bar{x} - \mu_0|}{\frac{\sigma}{\sqrt{n}}}))$$

- ג'ודל זה מכונה p והוא הסוג הראשון של השגיאה.
- מה ניתן לומר עליו?
- א. $0 \leq P \leq 1$.
- ב. חסר'יות
- ג. ככל שההבדל בין תוחלת המדגם לממוצע גדול יותר, הוא קטן יותר.

משמעותו של $P - value$: השיטה שצבעו הוא $P - value$ (כלומר, אם ההשערת H_0 נכונה, $P - value$ הוא שווה מאוז שנצפתה היא $p - Value$). $p - Value$ (כלומר, ככל שההפס נכונה) הסתברות ש- p יהיה או יותר קטן ש- $p - Value$ היא $p - Value$. $p - Value$ יכול להיות קטן, יכול להיות גדול.



צדדים ב- $P - value$:

- א. מבחן *right-tailed*: אם אנחנו חושדים מרأس (לפניהם שראינו את הנתונים) כי $\bar{x} < \mu_0$.
- ב. מבחן *left-tailed*: אם אנחנו חושדים מרأس (לפניהם שראינו את הנתונים) כי $\mu_0 > \bar{x}$.
- ג. מבחן דו צדי (*two-tailed*): בירית המחדל. חושים שהם שווים. נשים לב שההנוסחאות לחישובו ישתנו במבחן דו צדי.

סוג המבחן	ימני	双边	שמאלי
הערך הקritisטי -	$C = \mu + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$	$C^+ = \mu + Z_{a/2} \cdot \frac{\sigma}{\sqrt{n}}$ $C^- = \mu - Z_{a/2} \cdot \frac{\sigma}{\sqrt{n}}$	$C = \mu - Z_a \cdot \frac{\sigma}{\sqrt{n}}$
המבחן -	אם $E[X] < C$ נדחה H_0 אחרת מקבל.	צריך לקיים $C^- < E[X] < C^+$ או על מנת לקיים את H_0	אם $E[X] > C$ נדחה H_0 אחרת מקבל.

מה ה- $P - value$ לא אומר? הוא לא אומר אם קיבל או לדוחות את השערת האפס!!!

7.4 שלבים להחלטה אם לקבל או לדוחות את השערת האפס

- כדי להחליט אם לקבל או לדוחות את השערת האפס יש לבצע את הצעדים הבאים:
- א. להחליט לפני החישוב, על סף שאם $P - value$ – P יהיה קטן ממנה נדחה את השערת האפס (למשל, סיבים מקובלים הם $0.001, 0.05, 0.01$).
 - ב. לחשב את $P - value$ של המדגם.
 - 3. לדוחות את השערת האפס אם $P - value$ קטן מהסף שקבענו. לאחר שלב זה אין משמעות לוגדיו של $P - value$.

נניח שה- $P - value$ גדול מהסף שהוחלט מרأس. האם זה גורר ש- H_0 נכון? לא! זה רק אומר שאי אפשר לדוחות את השערת האפס. מדוע? אין מספיק נתונים, או שהמידע רועש מדי.

צורה נוספת לחושב על $P - value$: ההסתברות לקבל את המודגם (הנתונים) שקיבלנו, בהנחה שהשערת האפס היא נכונה.

הגדלה: אם החלטנו לדוחות את השערת האפס כיון H_0 היה נמוך מהסף שקבענו, נאמר שהשערת האפס נדחתה באופן מובהק סטטיסטי.

7.5 תיקון למבחנים מרובים

אם מבצעים מספיק מבחנים בסף $P - value$ נתון, הסבירות לקבל תוצאה קטנה מהסף עולה עם

מספק המבחןים. לשם כך נستخدم בתיקון *Bonferroni* - $P - value < \frac{\alpha}{n}$ באשר α הוא הסף למבחן בודד, ו n הוא מס' המבחןים. תיקון זה הוא שמרני, ישנים אלטרנטיביות. לעומת זאת, אם הסף הקודם היה α נאמר כי הסף החדש הינו:

$$\alpha' = \frac{\alpha}{n}$$

ואז, נדחה עבור מבחן ספציפי אם $P - value < \alpha'$

7.6 חישוב גודל המזגט הדירוש

קודם לכן ראיינו כי:

$$P(\mu_0) = 2(1 - \phi(\frac{|\bar{x} - \mu_0|}{\frac{\sigma}{\sqrt{n}}}))$$

ולכן -

$$n = \frac{\sigma^2}{(\bar{x} - \mu_0)^2} z_{\alpha}^2$$

וכך אפשר (תחת הנחות על ההתפלגות של X) לחשב כמה דוגמאות דרישות כדי לגלוות הבדל משמעותית סטטיסטי ברמת מובהקות נתונה.

7.7 מספר מקרים

עד כה - התיחסנו למצב שבו יש מבחן אחד שמשווה לערך תאוריti בודד. מה קורה באשר ישנים שתי מדגמים?

מדוע זה שונה? במצב זה, נדרש לנקוט בחשbon את הפרמטרים של שתי ההתפלגות ואת השכיחות היחסית של שתיהן.

עד כה - דנו בשאלת כיצד אפשר לשנות בשנייה מסווג 1: ההסתברות לדוחות את השערת האפס בטעות.

עוצמת המבחן: השם שנitin לFN – 1, ככלומר: המשלים של הסיכוי לשגיאה מסווג 2 – ההסתברות לקבל את השערת האפס בטעות (β). ככלומר, עוצמת המבחן היא ההסתברות לדוחות את השערת האפס כשהיא באמת שגואה.

עוצמת המבחן, היא ההסתברות לדוחות נכון את H_0 באשר H_1 נכון. אנו מחפשים, את המשלים של עוצמת המבחן. נראה כי, המשלים של עוצמת המבחן יהיה לקבל את H_0 בעוד H_1 נכון – זו בדיקת השגיאה מסווג 2.

كيف מחשבים את עוצמת המבחן? ראשית מחשבים את $\phi(\frac{C-\mu_1}{\sigma/\sqrt{n}})$ – בעוד $C = \beta$ והוא הערך הקרייטי (כתלי בבדיקה שקבע מתי נדחה את H_0 לפי רמת מובהקות (α), μ_1 הוא הממוצע האמיטי שלכארה נכוון תחת H_1 . ולאחר מכן, מחשבים את $1 - \phi(\frac{C-\mu_1}{\sigma/\sqrt{n}})$.

עוצמת המבחן תלויות: בבדיקה בו משטים, ברמת המובהקות, בנתונים (גודל המדים וכו'). גודל האפקט שהוא מנסה לאחות (נדבר על מושג זה בהמשך).

עוצמת מבחן גובהה תתקבל (באופן כללי) אם יש לנו – שונות נמוכה בנתונים, מדגם גדול, גודל אפקט גדול, דרישות נמוכות לרמת המובהקות.

במילים אחרות: עצמת המבחן היא ההסתברות לטעוס את ההבדל באשר הוא אכן קיים. ככלומר - לדוחות נכוןת H_0 אם H_1 נכון. אם יש לנו רופא שטוען H_0 התרופה עובדת ו- H_1 התרופה אינה עובדת. אז, עצמת המבחן היא ההסתברות לומר כי התרופה לא עובדת כאשר התרופה באמת לא עובדת. לכן, זה המשלים לשגיאה מס' 2: הסתברות שטוען כי H_0 נכון בעודה H_1 היא נכוןה. לכן, אם עצמת המבחן גבוהה, סיכוי טוב שנמצא את מה שאנו חפשים אם הוא באמת שם.

7.8 גודל האפקט (d של כהן)

מודדר C

$$d = \frac{\bar{x}_1 - \bar{x}_2}{s}$$

באשר s היא סטיית התקן. קיימים מدادים נוספים לגודל האפקט. הוא מותאר את העוצמה או הגודל של הקשר או ההבדל שיינו במדגם, באופן שאינו תלוי בגודל המדגם.
МОובקהות סטטיסטיות (P -value): אומרת לנו האם יש הבדל/אפקט (האם הוא אמיתי או מקרי).
גודל האפקט (d): אומר לנו כמה גודל ההבדל/האפקט.

7.9 סיכום חשוב

במציאות ישנים 2 מצבים בלבד. או H_0 נכון, לא קיים אפקט בכלל. או H_1 נכון, קיים אפקט כלשהו ומהו השתנה.
 אבל איןנו יודעים מהו המבחן האמתי. לפיכך,
 המבחן הסטטיסטי הוא למעשה קומשלתו.
 אם הטעאה קיצונית מסוימת \leftarrow יש אפקט \leftarrow דוחים את H_0
 אם הטעאה לא קיצונית \leftarrow אין אפקט \leftarrow לא דוחים את H_0
 אבל הuko זהה לא מושלם,uko הבדיקה יכול לטעות. נתבונן בשני עולמות מקבילים:

עולם ראשון - H_0 נכון (אין אפקט): בעולם זה, אם נרים את הניסוי הרבה פעמים ברוב הפעמים נקבל תוצאות רגילות ונגיד שאין אפקט. בחלק קטן מן הפעמים, ב-5% מהם (אם ללחנו $\alpha = 0.05$) אנחנו נקבל תוצאה קיצונית ונגיד בטעות "יש אפקט" בעודנו יודעים שאין אפקט שזו: זו בדיקת שגיאת מס' ראשון.
 עולם שני - H_1 נכון (יש אפקט): בעולם זה, אם נרים את הניסוי המון פעמים: ב- $\beta\%$ מהפעמים ($\beta = 1 - \text{power}$) אנחנו נקבל תוצאה קיצונית ונגיד: יש הבדל, בעוד אכן יש הבדל. זו בדיקת עצמת המבחן.
 במקרה $\beta < \alpha$, נקבל תוצאה לא מספקת קיצונית ונגיד בטעות: אין אפקט. זו בדיקת שגיאת מס' שני.

הערה חשובה: אם α קטן יותר, המשמעות היא שאזרה הדחיה הפך להיות קטן יותר - כיון שאנחנו דוחים את H_0 רק בערכים מאוד קיצוניים. ולכן, במקרה זה אזרה הקבלה הופך לגודל יותר.

7.10 מבחני השערות במדוגמים גדולים

מהו מדגם גדול? מקובל לומר שمدגם עם 30 או יותר דוגמאות הוא נחשב גדול. מדוע זה חשוב? מס' מדדים סטטיסטיים מתפלגים נורמלית במדגם גדול.
 לכן, אפשר להשתמש בחישובים שעשינו עד כה כדי לבדוק מובקהות סטטיסטית. אם המדגם קטן, יש לבצע תיקון למדדים.

מקרה מפורסם שכאלה: בהינתן n דוגמאות בלתי תלויות של משתנה נורמלי, נחשב:

$$t = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

הנתונים מותפלגים בהתפלגות t עם מס' דרגות חופש השווה $1 - n$. אם נרצה לחשב את התפלגות t זה באמצעות הנוסחה הבאה:

$$f(t) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{\pi v} \Gamma(\frac{v}{2})} (1 + \frac{t^2}{v})^{-\frac{v+1}{2}}$$

כאשר v הוא מס' דרגות החופש ו Γ היא פונקציית גאמא. לשמהנו, באשר v גודל התפלגות t שואפת להתפלגות נורמלית.

7.11 מבחני השערות

איך בודקים אם מה שוראינו במדגם שלנו הגיוני או קיצוני מדי לעומת מה שציפינו? נתcken את התובוצה שלנו למשתנה Z ונבדוק אם הוא נמצא בטוחה הסביר או שלא. תמיד אנחנו נניח כי H_0 נכונה, ונבדוק האם הנתונים סותרים זאת.

7.11.1 דוגמה ראשונה: ממוצעים

במצבי זה, יש לנו טענה על האוכלוסייה: הממוצע באוכלוסייה הינו \bar{x} . איננו יכולים לבדוק את כל האוכלוסייה. אז מה עושים? נדגם מוגן קטן ונבדוק מה הממוצע שם. השאלה שירצה לשאול, האם הממוצע שמצאנו באוכלוסייה סותר את הטענה המקורית או שלא? או במילים אחרות – האם ההבדל שמצאתי בין מה שטענו נבע בכלל רק מקרה בודד או רוש סטטיסטי או ממש משמעותי ואז הטענה המקורית אודות הממוצע הייתה נכונה.

- נסמן:
- א. \bar{x} ממוצע המדגם
 - ב. n תוחלת האוכלוסייה
 - ג. σ סטיית התקן של האוכלוסייה
 - ד. s גודל המדגם

המשתנה המתוקן יהיה $Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ והוא אומר "למעשה – כמה ייחדות רעש רוש מהתמוצע המוצע? וכעת, אם נרצה רמת מובהקות של α בטענה, אז לא נדחה את השערת האפס אם:

$$-\phi(\alpha) \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq \phi(\alpha)$$

וכן נדחה אותה אם הזבר אינו מתקין.

דוגמה לממוצעים. טענה: H_0 : הגובה הממוצע של גברים בישראל $= 175 \text{ cm}$. אנחנו נדגם $n = 100$ גברים אקרים. קיבל כי $177 = \bar{x}$ של האנשים שדגמנו. נניח כי $175 \neq 177$. מדוע זה קרה? אפשרויות ראשונה: זה קרה במקרה, אם הייתי דוגם עוד מאות היווני מקבל $= 173 \text{ cm}$. במקרה זה, הטענה המקורית עדין נכונה ולא דוחים את H_0 . אפשרות שנייה: ההבדל גדול מדי בשליל להיות מקרי. כיצד מחליטים? המבחן הסטטיסטי מוחלט. הוא שואל: מה הסיכוי לקבל ממוצע של 177 במדגם באשר הממוצע באוכלוסייה הינו 175 cm ? הוא קובע רמת מובהקות, ומחשב בהתאם.

7.11.2 דוגמה שנייה: הצלחות במדגם

כעת אנו במצב בו שיעור ההצלחות / כר / מסוימים באוכלוסייה הינו $p = \mu$. כמובן, μ הוא יחס ההצלחות במדגם (למשל, מס' הפעמים שניסוי הצלחה מותך כל הניסויים). דגמו מדגם, וקיבלנו שיעור $P = \mu_p$. כעת, נרצה לבדוק האם ההבדל בין P לבין μ הוא משמעותי?

כיוון שמדובר במשתנה בינומי, $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$. נרצה לתקן, המשטנה המתוקן הינו:

$$Z = \frac{\mu_p - \mu}{\sigma_p} = \frac{P - p}{\sqrt{\frac{p(1-p)}{n}}}$$

וכעת, אם נרצה רמת מובהקות של α : איזה נדחה את השערת האפס אם:

$$-\phi(\alpha) \leq \frac{P - p}{\sqrt{\frac{p(1-p)}{n}}} \leq \phi(\alpha)$$

וכן נדחה אותה אם הדבר אינו מתקיים.

דוגמה. נניח כי H_0 היא שבבחירה הקרוובות, $p = 0.4$ הוא אחוז התמייה במועד א' בבחירה. בסקר, שאלנו $n = 500$ אנשים ומתוכם $P = \frac{230}{500} = 0.46$ תומכים במועדם. נרצה לבדוק: האם $0.4 < 0.46$? נחשב את Z לפי הנוסחה ונקבל $Z = 2.74 > 1.96$. כעת, $2.74 > 1.96$ ברמת מובהקות של 5% ולכן דוחים את H_0 ברמת מובהקות 5%.

חשיבות: אם מתקיים $10 > np > n(1-p)$ איזה אפשר לומר כי פורפורציית המדגם מתפלגת נורמלית ככלומר ($\hat{p} \sim N(p, \frac{p(1-p)}{n})$) דהיינו $\hat{p} = p, \sigma^2 = \frac{p(1-p)}{n}$. כעת יוכל להשתמש ב- Z בשיביל לדעת את פונקציית הצפיפות המצטברת של משטנה בינומי.

7.11.3 דוגמה שלישיית: הפרשים בין ממוצעים

עד כה השווינו מדגם אחד לטענה תאורתית. כעת, נרצה להשוות בין שני מדגמים. כעת נרצה לענות על השאלה: האם יש הבדל משמעותי בין שתי קבועות? נניח כי קבועות בלתי תלויות זו בזו. כמובן: בדקנו שתי קבועות, אסיאТИים ולא אסיאТИים. הממוצע של האסיאТИים בגובהם הינו 174 ס"מ והממוצע של הלא אסיאТИים בגובהם הינו 177 ס"מ. אנו שואלים - האם ההפרש של 3 בנים היה מקרי, או שפשות הטענה אודות ממוצע הגבאים שלהם היא sama לא שווים. נפרמל,

יהיו שני ממוצעים μ_1, μ_2 . בהינתן כי H_0 הינה $\mu_1 = \mu_2$ והיא $\neq \mu_1$ (הממוצעים שונים). לכן, נמיר את המבחן לבודן של הבדל הממוצע מאפס, אותו אנחנו ידועים לחשב (מדוע? יוצרים משתנה חדש שהוא ההפרש, ובודקים את ההשוואה שלו אל אפס):

$$H_0 : \mu = \mu_1 - \mu_2 = 0$$

$$H_1 : \mu = \mu_1 - \mu_2 \neq 0$$

נסמן את הממוצעים בהתאם \bar{x}_1 ו- \bar{x}_2 ואת גדי המדגמים n_1, n_2 בהתאם. סטיית התקן של הפרשי המדגמים:

$$E_{S1-S2}^2 = E((X_1 - X_2)^2) =^* E(X_1)^2 - 2E(X_1 X_2) + E(X_2)^2 = E(X_1)^2 + E(X_2)^2$$

* שכן הגורם $-2X_1 X_2$ יוצר תוחלת 0 כיון שהגורמים בלתי תלויים. לכן, שגיאת התקן של ההפרש בין המדגמים הינה:

$$\sigma_{S1-S2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

המשתנה מתוקן של ההפרש בין הממוצעים הינו:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

מכאן, שתחת השערת האפס נקבל כי $0 = \mu_1 - \mu_2$ וכך:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

הערה חשובה. מואוד. אם $\mu_1 = \mu_2 + 2$ כהשערה האפס, אז כמובן ש $2 = \mu_1 - \mu_2$ ולא אפס ולכן

$$Z = \frac{\bar{x}_1 - \bar{x}_2 + 2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

כעת, אם יהיה מואוד גדול אליו ההפרש בנים כנראה לא היה אפס ונדחה את השערת האפס ולכן נאמץ את ההשערה האלטרנטיבית H_1 . במקרה של $\alpha = 0.05$ קיבל כי $Z_\alpha = 1.96$. נניח ו Z שלנו יצא $1.96 < Z = 60$, אז $60 > 1.96$! ההפרש עצום ולכן בהכרח נדחה את השערת האפס.

בأופן דומה, אם נסתכל על הפרש בין הצלחות בין קבוצות, כלומר האם שיעור ההצלחות בקבוצה 1 שונה משיעור ההצלחות בקבוצה 2? נבחן כי אם בוצעו n_1 ניסויים שהצליחו בהתאם x_1 , x_2 פעמים איז p , ומכאן:

$$\sigma_{P1-P2} = \sqrt{p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})}$$

ונקבל כי

$$Z = \frac{P_1 - P_2}{\sigma_{P1-P2}}$$

דוגמא. נניח כי $n_1 = 500$ גברים ראו פרסום $x_1 = 150$ לחצו עליה. אז $P_1 = \frac{150}{500} = 0.3$ וכן $n_2 = 600$ נשים ראו פרסום $x_2 = 120$ לחצו עליה אז $P_2 = \frac{120}{600} = 0.2$. כיצד נבדוק את ההשערות? השערת האפס הינה $P_1 = P_2$ והאלטרנטיבית היא שهما שונים. לפיכך, נחשב את Z לפי הנוסחה מלמטה ונקבל כי $Z = 3.85 > 1.96$! נדחה את H_0 .

7.11.4 מבחנים של זוגות

עד כה הדגימות היו בלתי תלויות זו בזו. ומה אם הדגימות כן תלויות זו בזו? נניח שיש לנו זוגות של מדידות $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - תלויות זו בזו. למשל, משקל לפני ואחרי דיאטה. כמובן שיש תלות. לכל זוג כזה נחשב את ההפרש המתאים:

$$D_i = x_i - y_i$$

כעת בידינו n דגימות: D_1, D_2, \dots, D_n של ההפרשים בין התוצאות. נבחן כי כעת הם בלתי תלויות. (**תמיד להזכיר בדגםאות עם ההבעות לטראם.** (x_1, y_1) זה ההבעות לטראם בפלורידה ב-2016, (x_2, y_2) זה ההבעות לטראם ב-2020 בהתאם). וכן אם נסתכל על ההפרש בין ההתאמות, קיבל סדרה חדשה של 50 המדיניות בלבד שלא תלויות זו בזו. **כלומר סדרה של נתונים בלתי תלויים.** לסדרת ההפרשים ניתן לחשב תוחלת וסטיית תקן. כעת,

$$H_0 : \mu_x = \mu_y$$

$$H_1 : \mu_x \neq \mu_y$$

כעת אם נמיר לדפוס D נקבל כי:

$$H_0 : E(D) = \mu_x - \mu_y = 0$$

$$H_1 : E(D) = \mu_x - \mu_y \neq 0$$

כעת המשתנה המתוקן יהיה:

$$t = \frac{E(D) - 0}{S(D) / \sqrt{n}}$$

שכן כעת התוחלת היא 0 מתחת השערת האפס (אותה אנו מניחים ב מבחני הרשענות). נשים לב כי נחתת היסוד של המודל היא שההפרשים אכן מתפלגים נורמליים.
הערה. נבחן כי בכוונה סימנו t , Z זה כאשר אנו יודעים את σ סטיית התקן האמיטית ו t זה באשר אנחנו מעריכים את σ מתוך המדגם S . אם $30 < n$ נוכל להשתמש ב t , ואם $n \leq 30$ ניהיה t .
 חייבים להשתמש בטבלת.

لسיכום: מתי נשתמש ב מבחן זוגות? נשתמש כאשר אנחנו במצב של לפני או אחרי - מודדים את אותו הדבר פעמיים, ולא להשתמש ב מבחן זוגות כאשר הזוגות באמת בלתי תלויות כמו נשים ובברים, אסיאתים ולא אסיאתים וכו'.

7.11.5 **מבוחן אפרמטריים: χ^2**

פרמטרים הם מאפיינים של התפלגות. למשל, ממוצע, סטיית תקן, צורות התפלגות כמו נורמלית או פואסון.

עד כה עבדנו עם מבחןים פרמטריים, הנחנו הנקודות על הנתונים - הנתונים מתפלגים נורמלית, התוחלת היא μ , סטיית התקן היא σ וכדומה. מבחן Z זה מבחן פרמטריים. אם כן, אלו מבחנים שרגילים הרבה יותר לנtíונים חריגים, דורשים שהנתונים יהיו נורמליים ולא עובדים טוב על נתונים קטגוריאליים.

מבחןים אפרטוריים הם מבחנים שאינם מוחים הנחות על התפלגיות הנתונים. לכן: פחות רגילים לנוטונים יוצאי דופן, אין צורך לבדוק את התפלגות הנתונים והם עובדים הן על נתונים אורדינליים והן על נתונים קטגוריאליים. מה החסרונות? הם חזקים פחות מ מבחנים פרטוריים - דורשים מוגדים יותר.

מבחן² הוא מבחן אפרטורי שיעונה על השאלה הבאה: האם ההתקפלות שנייה רואה בנתונים תואמת להתקפלות צפיפיתיה לה? למשל: זריקת קוביה - יש לך קובייה ואותה זורק אתה 600 פעמים. אם הקובייה הוגנת, אתה מזכה לראות במספר 100 כל מס' 100 פעמים. אם בפועל קיבלתך רק 107 עבור 1 ו-931 עבור 2 למשל, האם הבדל זה קרה בנסיבות מקרים רבים או בכוונה והקוביה אינה הוגנת?

הרעיון זהה, נמדד ככמה רוחקה ההתפלגות שראינו מההתפלגות שציפינו. לשם כך, נתבונן בנוסחה הבאה:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

הפרש יצא 20 אך קיבלנו 20 וציפינו 40, זה לא אותו דבר כמו שציפינו ל-120 וקיבלו נ-100.

לוקחים דוגמה של קוביה למשל, עם נתונים ידועים. מוחשבים ומתקבלים כי $\chi^2 = 2.58$. מס' דרגות החופש הינו $k - 1$, כיוון שישנם 6 ערכים (אפריים של מספרים בקוביה) איז מס' דרגות החופש הינו 5. שכן, אם קבענו 5 ערכים עבור כמה הטילות יצא בהם 1, 2, 3, 4, 5 את $\alpha = 0.05$ (רמת ודאות) והוא 600 ידועה התזacha האחרונה.icut, מוחשבים בטבלת χ^2 את $\chi^2 = 2.58 < 11.07 = C$ (ערך קritis). כיוון שאצלנו לא דוחים את H_0 הקוביה הינה הוגנת.

באופן כללי עבור מבחון χ^2 : אם נקבל $\chi^2 > C$ אז לא נדחה את H_0 ואם נקבל $\chi^2 < C$ כן נדחה את H_0 במבחן החשיבות.

וכן, אם χ^2 קטן אין זה קרוב למה שציפינו, יכול להיות רעש ורגיל ולכון לא דוחים. אם χ^2 גדול אין זה רחוק ממה שציפינו לא יותר מאשר רק רעש, ולכון דוחים.

הערה חשובה. נבחין כי χ^2 בהכרח חיובי, שכן אם $0 = \chi^2$ זו התאמה מושלמת, בדיקות מה שרצינו. בכונונה מדובר ב χ^2 שהערך יהיה תמיד חיובי. הערך הקרייטי שהרי הוא קו הגבול הוא יקבע האם נדחה או לא נדחה.

7.11.6 מדידת הפרשים א-פרמטרית במדגמים בלתי תלויים: מבחן Whitney U TEST

מדוע אנחנו זוקקים ל מבחן נוסף? עד כה, בשביל להשווות שתי קבוצות השווינו ממצאים עם מבחן t .

אבל מבחן t דוחה הhipothesis נורמלית, מקרים גדולים ונtiny נריצים. מה אם הנתונים לא נורמליים? המודגש קטן? יש ערכים קיצוניים? ישנו נתונים אוריינטליים (דירוג, לא מספרים)? לשם כך משתמש בבחן הא-פרמטרי הבא שלא דוחה הנחות על ההתפלגות או הפרמטרים.

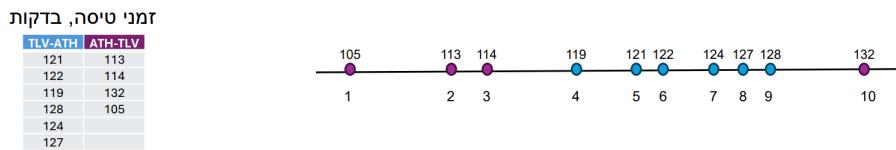
המודגש בודק האם החזיניות של שני התפלגות זהים. משווים את החזיניות של שתי ההתפלגות. נתנות דגימות x, y מהתפלגות אוריינטליות רציפות F_x, F_y בהתאם. כיוון שההתפלגות רציפות $P(x = y) = 0$:

$$H_0 : P(x > y) = P(y > x) = \frac{1}{2}$$

$$H_1 : P(x > y) \neq P(y > x)$$

כלומר, אנחנו בודקים האם החזיניות של התפלגות שווים. אם הם שווים אז ההסתברות למשווה מעליו או מתחתיו זהה.

נתבונן בדוגמה: האם הטיסות מישראל לאטונה יותר זמן מתisosota מאטונה לישראל?



המבחן מציע את הרעיון הבא: נסדר את הנתונים על קו ישר" באשר נזכיר לכל נקודה מהיכן היא הגיעה - מאיו דגימה וכן נסדרם בסדר עולה. נבחן כי אם H_0 נכון (אין הבדל) אז הם צריכים להיות מעורבים (כי החזיניות שלהם שוים), ואם H_1 נכון אז הערכים יהיו מקובצים בקבוצות. וכן, נחשב את

$$R_1 = 1 + 2 + 3 + 10 = 16$$

$$R_2 = 4 + 5 + 6 + 7 + 8 + 9 = 39$$

כעת, הרעיון הוא שם אין הבדל, אז הדירוגים צריכים להיות מעורבים באקראי ואם יש הבדל הדירוגים יהיה מקובץ. נבחר את סך המיקום הקטן יותר ונשאול, היכן הוא ביחס לכל הסדרים האפשריים? נסתכל על $R_1 = 16$, זה סכום די קטן ויש לפחות 1 ערך בערך קטנים. ישנו $\binom{10}{4}$ סדרים אפשריים (בוחרים 4 לסטיגמים, וכל השאר כבר יסתדרו בהתאם) לחלק את הדירוגים. וicut נסתכל על סכום הסדרים האפשריים. נראה היכן עומד R_1 ביחס לסדרים. במקרה שלנו, אם נחשב (המחשב יחס) קיבל כי רק 13% מהפרומות יותר קטנות (כלומר, רק 13% מהפרומות מוגבלות מ-16). ולכן היכן עומד R_1 מ-16. ולכן זה לא מובהק סטטיסטי - זה לא דוחה את השערת האפס. אם היינו מקבלים ממש הפרומות יותר קטנות יותר הוא קטן מ-5% אז זה היה מובהק סטטיסטי.

ובאופן כללי -

מדוע אנחנו צריכים משתנה U ? R הוא תלוי בגודל המדגמים. ולכן, אנחנו מונרמלים את R בהינתן גודל המדגמים. שוב, R_y זה סכום הדירוגים של הקבוצה y ו n_y זה מס' התכפיות בקבוצה זו. אנחנו נNORMALIZE אותו על ידי כך שנוריד ממנו את הפרטוטזיה המינימלית האפשרית: זו שנקח בה את המיקומים n , $1, 2, 3, \dots$ (הסכום יהיה הילך קטן).

$$U_y = \sum_{j=1}^{n_y} (R_{y,j} - j) = R_y - \frac{n_y(n_y + 1)}{2}$$

כעת, U יאמר לנו כמה אנחנו מעלה המינימום. אם נחזיר לדוגמה הקודמת, מותקים $U_1 = 6 - \frac{4 \times 5}{2} = 16 - 10 = 6$, כלומר אנחנו 6 נקודות מעלה המינימום. נבחין כי U המינימלי הוא זה שהסדר בו הוא המינימום ולכן $U = 0$. אם U ביןוני - הקבוצה מעורבתת ואם U גדול - הקבוצה מוקובצת בחלק העליון.

כל שהמדד הזה יותר גדול, ההבדל פחות מובהק.

ישנה טבלת U , לא משתמש בה. תמיד יכתבו לנו שמדובר במספיק גדול ולבן משתמש בקריבוב הבא: נבחן כי בדוגמא קטע יכולנו לחשב את מספר האפשרויות לסדר, אך במקרים עם n גדול זה מספר אסטרטוני. לפיכך, כשהמדובר במספיק גדולים $20 - 10 > n_1, n_2 > 10$ נשתמש ב מבחן Z וגיל באשר החישוב יהיה כך:

$$\begin{aligned} m_U &= \frac{n_1 n_2}{2} \text{ ויחושב כך:} \\ \sigma_U &= \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \\ Z &= \frac{U - m_U}{\sigma_U} \end{aligned}$$

דוגמא. האם גברים גבוהים יותר מאשר נשים? 50 גברים ו 60 נשים. אחרי סידור כל 50 ו 60 האנשים יחד קיבלו Ci $R_1 = 3200$ (סכום המיקומים) של הגברים. ראשית נחשב את U :

$$U = 3200 - \frac{50 \times 51}{2} = 1925$$

$$m_U = \frac{50 \times 51}{2} = 1500$$

$$\sigma_U = \sqrt{\frac{50 \times 51 \times (50 + 51 + 1)}{12}} = 166.58$$

$$Z = \frac{1925 - 1500}{166.58} = 2.55$$

הערך הקritisטי באשר $\alpha = 0.05$ הינו $Z = 1.96$, ב מבחן דו צדדי נדחה את H_0 כי $2.55 > 1.96$. כלומר: גברים גבוהים יותר מאשר נשים. נבחן כי אינטואטיבית אכן הגברים גבוהים יותר שכן אם זה לא היה הממוצע הממוצע לגבר היה $\frac{\sum_{i=1}^{110} i}{110} = 55.5$ אך זה גדול מהממוצע - ככלומר הגברים מקבלים

דירוג גובה יותר מה ממוצע הכללי.

הערה חשובה. אם יש שני ערכים זרים הם יקבלו את אותו ערך דירוג. הערכה שנייה. علينا לחשב ערכי U לכל הקבוצות, ולחזור את המינימלי בניהם ואיתו לcliffe לטבלה.

שלבי המבחן:

- שלב ראשון: סדר אותם לפי סדר עולה.
- חשב את הראנקים R_i
- הגדר את U_i המתאים
- הגדר $U = \min\{U_1, U_2\}$
- בדוק בטבלה עבור n_1, n_2, α . אם $U \leq C$ נדחה את השערת האפס ואם $U > C$ לא נדחה אותה.

7.11.7 מדידת הפרשים א-פרמטרית בדגם מזוג (מבחן Wilcoxon Singed-Rank Test)

מדוע אנחנו צריכים מבחן נוסף? במקרים בלתי תלויות ישנו מבחן פרמטרי של Mann – Whitney – U . במקרים תלויות: יש לנו את מבחן t לזוגות וכעת נלמד מבחן א-פרמטרי לזוגות.

נשתמש במבחן זה כאשר אוטם אנשים נבדדו פעמיים (לפni, אחריו), זוגות מותאמים (תאומים, אותו אדם), הנתונים לא נורמליים או יש חריגות, מדגם קטן. כמובן, נניח שיש שני מדגמים מזוגיים (x_i, y_i). נרצה לבדוק אם החיצון של x_i שונה מהחיצון של y_i .

לדוגמה, האם מרוץן בוסטון בשנת 2024 היה מהיר יותר ממרטון בוסטון בשנת 2023?

Runner	2023	2024	Difference	Ordered
Evans Chebet	125.9	127.4	1.5	0.2 (-)
Albert Korir	128.0	127.8	-0.2	0.7 (-)
Talbi Zouhair	128.6	130.8	2.2	1.0
Hellen Obiri	141.6	142.6	1.0	1.2
Ababel Yeshaneh	144.0	146.2	2.2	1.5
Emma Bates	142.2	147.2	5.1	2.2
Hiwot Gebremaryam	144.5	145.3	0.8	2.2
Matthew McDonald	130.3	141.9	11.6	5.1
Isaac Mpofu	134.1	128.3	-5.8	5.8 (-)
Cj Albertson	130.6	129.9	-0.7	11.6

מדובר בזוגות – זה אוטם אנשים בדיק, שרצים כל אחד פעמיים: פעם אחת בכל שנה. אי אפשר לטעון שאלה זוגות בלתי תלויים – יש תלות בין כל שנים. נסמן:

$$H_0 : m_{2023} = m_{2024}, H_1 : m_{2023} \neq m_{2024}$$

בשלב הראשון של המבחן – מחשבים הפרשים. חיובי + אומר כי רץ איטי יותר ב-2024 ושלילי – אומר כי רץ מהיר יותר ב-2024. מסדרים את הערכים לפי גודל ההפרש – ללא סימן. לאחר מכן, לכל ערך מוחזירים את הסימן שלו ומתקבלים סידור שלהם באופן מדורג. מחשבים את w^+ ו- w^- : שתי קבוצות המייצגות בהתאם את האינדקסים המדורגים של האיברים סימנים חיובי, ואייברים שיש מהם שלילי. מתחבוננים ב- w^+, w^- : $W = \min\{w^+, w^-\}$ – נרצה לדעת את המיעוט המשמעותי.

אם אין הבדל - אינטואטיבית נרצה כי מחצי מההפרשים חיוביים ומחצית שליליים ככלומר
 $w^+ = w^- = \frac{\sum_{i=1}^n i}{2}$

הולכים אל הטבלה: Wilcoxon Singed-Rank, מתבוננים במס' הזוגות $n = 10$ וכן $\alpha = 0.05$
 ומתקבלים ערך קרייטי $C = 8$. קיבלנו כי $W = 12 > 8 = C$. ולכן לא דוחים את H_0 . **למען האינטואיציה - המבחן בודק האם ההפרשים מעורבבים באופן סימטרי.**

אם כן: לא נרצה את הטבלה. במדגמים גדולים ישנו קירוב של W כז' שיתפלג נורמלית:

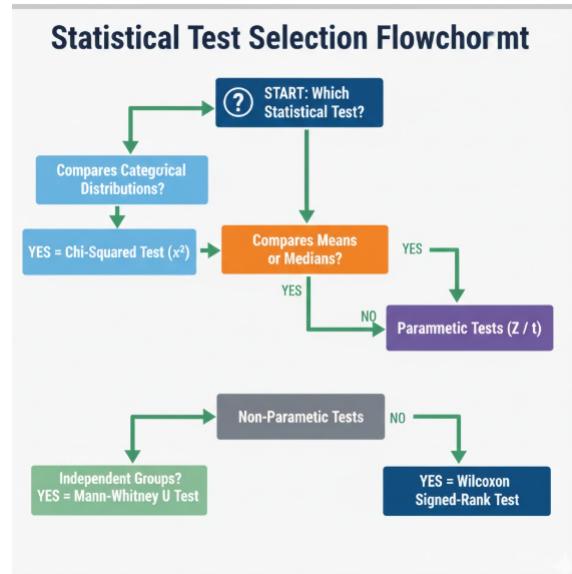
$$\mu_W = \frac{n(n+1)}{4}, \sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}, Z = \frac{W - \mu_W}{\sigma_W}$$

לסיכום, שלבי המבחן:

- א. חשב את ההפרשים (X_i, Y_i) $D_i = X_i - Y_i$ לכל זוג i .
- ב. הסר אפסים: אם $D_i = 0$ עדכן את n בהתאם והוציא את הזוג מהמדגם.
- ג. לכל זוג קח את ההפרש $|D_i|$ בערך מוחלט.
- ד. סדר את $|D_i|$ בסדר מהקטן לגדול.
- ה. תן דירוג $n, \dots, 1$ לכל אחד מהערכים.
- ו. החזר סימנים - תן לכל דירוג את הסימן המקורי.
- ז. חשב סכומים: W^+ , W^- שמייצגים את סכום החיוביים (דרגותיהם) וסכום הדרגות השליליות בהתאם. הגדר $W = \min\{w^+, w^-\}$
- ח. אם $n \leq 20$: השווה את W לערך הקרייטי (n, α) בטבלה Wilcoxon. H_0 דוחים את $W < C$.
 .1 אם $W < C$
 .2 אם $W \geq C$ לא דוחים את H_0
- ט. אם $n > 20$ (מדגם גדול) חשב את $Z = \frac{W - \mu_W}{\sigma_W}$
- י. השווה לערך Z עם רמת ודאות α . אם בתחום - אל תדחה, אחרת תדחה את השערת האפס.

7.12 סיכום מבחני השערות

באיזה מבחן להשתמש?



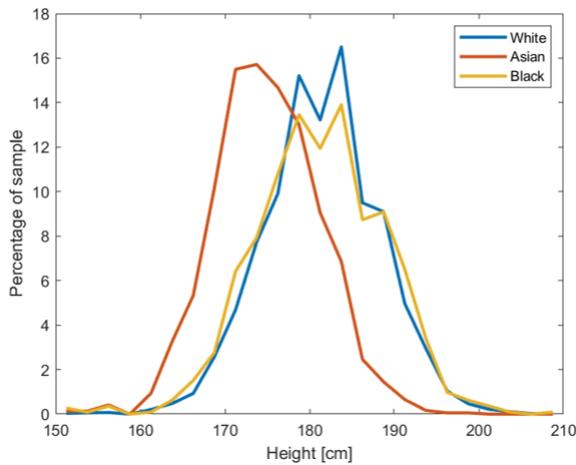
הרצאה 9 : 8 (Analysis of Variance) Anova

עד כה, דיברנו על בדיקת השערות במשתנה אחד (או שניים). כתע נדבר על בדיקת השערות באשר יש לנו יותר מ-2 קבוצות. בתחילת, נדבר כיצד נעשה זאת במשתנה אחת ובהמשך נרחיב ליותר משתנות אחת. (כלומר, נדבר על בדיקת השערות של משתנה אחד ביותר מ-2 קבוצות, ובהמשך נרחיב לבדיקת השערות של כמה משתנים ביותר מ-2 קבוצות).

עד כה ראיינו כיצד להשוות את הממוצעים של שתי קבוצות. מה נעשה אם יש יותר משתי קבוצות?

לדוגמא: האם גברים ממוצא אסיאתי שונים בגובהם מגברים אחרים? נניח, כי בתנאים יונק הרבה קבוצות נוספת: אנשי לבנים, אסיאטים, שחורים, יהודים, אינדיינים וכו' הלאה. כתע - ניתן לחלק את התנאים ליותר מ-2 קבוצות ולשאול שאלה אחרת: האם גברים שונים בגובהם כתלות בקבוצת האתנית שאליה הם משתייכים?

תבונן בהתפלגות הנכפית של הגברים:



זה נראה כי התפלגות הלבנים והשחורים יחסית זהה - אך התפלגות האסיאטים שונה מהם.
 ===== טעות! מה זה נראה? כבר אמרנו: צריך לבצע בדיקת השערות באופן מתמטי.

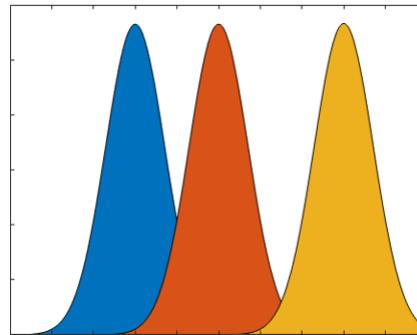
8.1 הנחות יסוד לבחן אנוּבָה (בסיסי)

- מטרה של המבחן:** להשוות בין יותר מ-2 קבוצות.
 א. ההתפלגות של כל אחת מן הקבוצות היא נורמלית. (כלומר ההתפלגות של כל משתנה שמעניין אותנו בכל קבוצה היא נורמלית, גאוסיאנית).
 ב. כל הדגימות נדגמו באופן אקראי ובלתי תלוי (אין קשר בין הדגימות בכלל)
 ג. לכל הקבוצות ישנה שונות שונה.
 ד. הגורם המבדיל (*factor*) בין הקבוצות הוא קטגוריאלי - למשל: גז שהם מגעים מהם, לא משתנה רציף כלשהו!
 ה. המשתנה שנמדד הוא רציף.

בහמשך, ננסה ליותר על חלק מההנחות. אך: עם ההנחות חיינו יהיו פשטוטים הרבה יותר.

8.2 הרעיון המרכזי של מבחן אנוּבָה

הרעיון המרכזי של מבחן אנוּבָה הוא להשווות את הפיזור של הנקודות מכל קבוצה ולהסתכל האם הפיזור בתוך הקבוצה הוא יותר הדוק מפיזור הנקודות בין הקבוצות. ככלומר: המבחן שלנו ישווה את פיזור הנקודות בתוך כל קבוצה לבין פיזור הנקודות בין הקבוצות - אם הפיזור בין הקבוצות גדול הרבה יותר מאשר הפיזור בתוכן נאמר שהמוצע שלן שונה באופן משמעותי.
 לדוגמה: עם הדוגמה על הגברים האסיאטים, הנקודות בתוך הקבוצה של הגברים הלבנים יותר דומות אחת לשנייה בפיזור לעומת המבוקש בין קבוצת הגברים הלבנים לשחורים.
 בצהורה גרפית: נרצה שהגואיסיאנס יהיה יותר צפופים אחד מוהני ופחות מעורבבים אחד בין השני - ואם זה יהיה נכון: יוכל לומר כי ישנה הפרדה לפי הקטגוריה שהקבוצות מגיעות מהן:



8.3 הגדרה פורמלית

הגדרה - **בחן אנוּבָה למשתנה אחד:** נסמן ב- μ_i את התוחלת של המשתנה i . ההשערות:

$$H_0 : \mu_j = \mu_k \forall j, k$$

$$H_1 : \mu_j \neq \mu_k (\text{for - some } j, k)$$

ההשערות הן: השערת האפס היא שלא משנה איזה 2 קבוצות נבחר, התוחלת של המשטנה הn'ל תהיה זהה בין הקבוצות. ההנחה האלטנרטיבית אומרת כי מספיק שקיים זוג אחד שהתוחלת בניהם שונה. (או הנחה מאד מחרמיה! מספיק שזוג אחד יראה לנו שההנחה לא סבירה בשליל את השערת האפס).

אמרנו כי ANOVA בודק הבדל בין פיזורים - בין שונות, אבל איך זה מסתדר עם העובדה שאנו מסתכלים על ממוצעים ותוחלות? זהו היופי שלanova - אנחנו נסתכל על פיזור הנקודות על variance וממנו אנחנו נסיק על איך התוחלות מתנהגות.

נתונות n דגימות מ q קבוצות: $n = n_1 + \dots + n_q$.
 תחת הנחות היסוד שלנו, נוכל לרשום את הערך של כל נקודה במדגם כ $y_{i,j} = \mu_j + \epsilon_{i,j}$ באשר $\epsilon_{i,j} \in N(0, \sigma)$ (סטטיסטית התקן אינה ידועה - היא קיימת, לא ידועה, בהמשך נסביר או נשער אותה).
 מדוע נוכל? אנחנו רושמים כל נקודה במדגם כסכום של תוחלת של הקבוצה של האיבר הספציפי ועוד איישחו רעש - סטטיטה סביב התוחלת של הקבוצה שלו.

למשל, עבור גברים אסייתים: הגובה הממוצע של גברים אסייאטים + סטטיטה ימינה או שמאליה.

תחת ההנחה כי המודל הn'ל של המשואה הוא נכון, ניתן לבצע כל מיני חישובים. למשל:
 א. ממוצע הנתונים בתוך הקבוצה הוא: $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j} + \epsilon_{i,j}$ באשר $1 \leq j \leq q$.
 ב. הממוצע של כל הנקודות הוא: $\bar{y} = \frac{1}{n} \sum_{j=1}^q \sum_{i=1}^{n_j} y_{i,j} = \frac{1}{n} \sum_{j=1}^q n_j \bar{y}_j$.

במבחןanova אנחנו רוצים לבדוק את פיזור הנקודות סביב הממוצע. לכן, בתחילת נגיד את פיזור הנקודות סביב הממוצע הכללי:

$$SS_{TOTAL} = \sum_{j=1}^q \sum_{i=1}^{n_j} (y_{i,j} - \bar{y})^2$$

זהו הפיזור הכללי של הנקודות במדגם. נוכל לפרק את הביטוי הפנימי:

$$\sum_{i=1}^{n_j} (y_{i,j} - \bar{y})^2 =_* \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2 + n_j (\bar{y}_j - \bar{y})^2 = (n_j - 1)s_j^2 + n_j (\bar{y}_j - \bar{y})^2$$

הטבר עכו * (זה לא היה כהרצאה, אוו הוסיף לעמו הגדנה):

$$\sum_{i=1}^{n_j} (y_{i,j} - \bar{y})^2 = \sum_{i=1}^{n_j} y_{i,j}^2 - 2y_{i,j} + \bar{y}^2 = \sum_{i=1}^{n_j} y_{i,j}^2 - 2\bar{y} \sum_{i=1}^{n_j} y_{i,j} + n_j \bar{y}^2$$

וינו הצד השני:

$$\sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2 + n_j (\bar{y}_j - \bar{y})^2 = \sum_{i=1}^{n_j} [y_{i,j}^2 - 2y_{i,j}\bar{y}_j + \bar{y}_j^2] + n_j \bar{y}_j^2 - 2n_j \bar{y}_j \bar{y} + n_j \bar{y}^2 =$$

$$= \sum_{i=1}^{n_j} y_{i,j}^2 - 2 \sum_{i=1}^{n_j} y_{i,j} \bar{y}_j + n_j \bar{y}_j^2 + n_j \bar{y}_j^2 - 2n_j \bar{y}_j \bar{y} + n_j \bar{y}^2 =$$

$$\sum_{i=1}^{n_j} y_{i,j}^2 + n_j \bar{y}^2 - 2 \sum_{i=1}^{n_j} y_{i,j} \bar{y} + 2n_j \bar{y}^2 - 2n_j \bar{y} \bar{y}$$

כעתعلילנו להסביר מזוע: ואז נוכיה
 $-2\bar{y} \sum_{i=1}^{n_j} y_{i,j} + 2n_j \bar{y}^2 - 2n_j \bar{y} \bar{y} = -2\bar{y} \sum_{i=1}^{n_j} y_{i,j}$ את השוויון.
 אם כן מתקיים $\sum_{i=1}^{n_j} y_{i,j} = n_j \bar{y}$ וכן:

$$-2n_j \bar{y}^2 + 2n_j \bar{y}^2 - 2n_j \bar{y} \bar{y} = -2n_j \bar{y} \bar{y}$$

$$\text{וכי } -2\bar{y} \sum_{i=1}^{n_j} y_{i,j} = -2n_j \bar{y} \bar{y} \text{ כנדרש.}$$

ושוב, מתקיים כי s_j הוא משערך השונות של הקבוצה j .
 קיבלו כי:

$$SS_{TOTAL} = \sum_{j=1}^q [(n_j - 1)s_j^2 + n_j(\bar{y}_j - \bar{y})^2]$$

ומכאן: סכום הריבועים הכללי נגזר מני דברים:

- א. הפיוררים של הקבוצות סביב הממוצע של הקבוצה
 - ב. המרחק בין הממוצע של כל קבוצה מהממוצע הכללי (ימין).
- נחלק את הסכום הנ"ל לשניים:

$$SS_{TOTAL} = \sum_{j=1}^q [(n_j - 1)s_j^2] + \sum_{j=1}^q n_j(\bar{y}_j - \bar{y})^2$$

נאמר כי:
 הפיזור סביב הממוצע של כל קבוצה:

$$SS_{Residual} = \sum_{j=1}^q \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2 = \sum_{j=1}^q [(n_j - 1)s_j^2]$$

הפיזור (המרחק) בין הממוצע של כל קבוצה מהממוצע הכללי מושקל בגודל הקבוצה:

$$SS_{Between} = \sum_{j=1}^q n_j(\bar{y}_j - \bar{y})^2$$

ובמילים אחרות:

$$SS_{TOTAL} = SS_{Residual} + SS_{Between}$$

פייזור הנקודות סביב הממוצע הכללי נובע מ:

1. פייזור הנקודות סביב הממוצע של כל קבוצה.

2. פייזור הנקודות של ממוצע כל קבוצה סביב הממוצע הכללי.

בשביל שנוכל להפעיל סטטיסטט (כמו Z או t) עלינו לדעת את מס' דרגות החופש:

א. מקור השונות - בין הקבוצות: ישנו $q - 1$ דרגות חופש. סכום הריבועים הוא

$$S_{\text{Between}}^2 = \frac{SS_{\text{between}}}{q-1}$$

ב. מקור השונות - בתוך הקבוצות $q - n$. סכום הריבועים הינו SS_{Residual} וממוצע סכום

$$S_{\text{Residual}}^2 = \frac{SS_{\text{Residual}}}{n-q}$$

ג.סה"כ שנות: $1 - n$ דרגות חופש, סכום הריבועים הוא SS_{TOTAL}

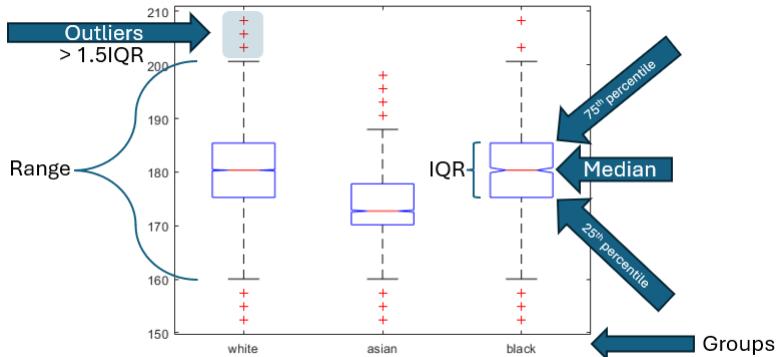
כפי שאמרנו, אנו רוצים לבדוק את היחס בין פייזור הנקודות בתוך הקבוצות לבין פייזור הנקודות בין הקבוצות. לכן, המדר הסטטיסטי שנבחן הוא:

$$F = \frac{S_{\text{Between}}^2}{S_{\text{Residual}}^2} = \frac{\frac{SS_{\text{between}}}{q-1}}{\frac{SS_{\text{Residual}}}{n-q}} = \frac{\frac{\sum_{j=1}^q n_j (\bar{y}_j - \bar{y})^2}{q-1}}{\frac{\sum_{j=1}^q [(n_j - 1)s_j^2]}{n-q}}$$

כל ש F גדול, אז קל יותר לדוחות את השערת האפס. מדוע? אם F גדול, אז המונה גדול, כלומר S_{Between}^2 גדול והממוצעים רחוקים זה מהה - שכן הערך במונה גדול (ולכן השערת האפס לא נcona) וכן זמנית, המונה קטן, כלומר S_{Residual}^2 קטן ולכן הגausיאנים נהיים צפופים (s_j קטן)

לאחר שיחסבנו את F , הולפים לטבלה של התפלגות F עם $n - q, q - 1$ ורמת המובהקות של רלוונטיות לנו נמצא שם ערך x . אם $x > F$ עבר את הגבול: נדחה את השערת האפס, אחרת לא ניתן לדוחות את השערת האפס.

לכל מבחן אנובה ניתן לקבל דיאגרמת קופסה. עבור הדוגמה של הגובה של הגברים נקבל:



הקו האדום מצין את החציוון, הקופסה את הטווח של $75\% - 25\%$ זה יקרא IQR . הטווח יהיה $1.5IQR$ (קונבנצייה). מדוע זה עוזר? במקרה של התפלגות ולומר כמה הם שונים - וכל חסתקכל על דיאגרמה זו ולהסתכל על הנתונים. לאחר שללנו את השערת האפס: אנחנו יודעים שלפחות יש זוג אחד שונה. עבור הדוגמה שלנו אם נרים אנובה נקבל שאכן השערת האפס לא נcona ממש - ואם נשמש בדיאגרמת הקופסה, יוכל לראות שזוגות שונים הם האסיאטים והלבנים.

- כלומר: קיבלו תוצאות בטן שמדובר באסיאתים מול הלבנים או האסיאתים והשחורים או שניהם
 - אך שוב: זה כמובן לא ניתן לומר באופן פורמלי. علينا לומר מתמטית מי שונה.

8.4 כיצד מתגברים על ההנחה שהנתונים מתפלגים נורמלית?

8.4.1 איך מתגברים על ההנחה שהנתונים מתפלגים נורמלית?

מציעים מבחן Kruskal wallis (מבחן Rank Sum) . מה עשו בעבר במקרה של קבוצות הנקה דבר על הקלט - ערכנו למבחרים אפרמטריים. הסתכלנו על הסדר וזה מה שקרה כאן:

- נחלף את הערכים בסדר שלהם במדגם.
- הסדר הממוצע בתוך קבוצה הוא $\bar{r}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} r_{j,i}$ (יחליף את הממוצע של המשתנה בתוך הקבוצה)
- הגסדר הממוצע של כל המדגם הוא $\bar{\bar{r}} = \frac{1}{n}(1 + 2 + \dots + n) = \frac{n+1}{2}$ (שהרי r מצין את הסדר בקבוצה והסדרים הם $n, \dots, 1$).
- נתבונן בגודל הסטטיסטי הבא:

$$H = \frac{SSR_{BETWEEN}}{SSR_{TOTAL} \setminus (n-1)} = \frac{\sum_{i=1}^q n_i (\bar{r}_i - \bar{\bar{r}})^2}{\sum_{i=1}^q \sum_{j=1}^{n_i} (r_{i,j} - \bar{\bar{r}})^2 \setminus (n-1)}$$

(דומה לאנובה, לא מדוייק).

מדוע $SSR_{Residual}$ לא במכנה? כיון שלא נרצה להניח דבר על ההתפלגות בתוך הקבוצות. אם יש לפחות 5 דגימות בכל קבוצה, אז נוכל לומר כי $H \sim \chi^2_{q-1}$ מתפלג כמו χ^2 עם $q-1$ דרגות חופש).

אם אין אף ערך שחוור על עצמו (ולכן כל אחד מקבל מקום ייחודי משלו בדירוג הסדר) אז מתקיים $SSR_{TOTAL} = \frac{(n-1)n(n+1)}{12}$

$$H = \frac{12}{n(n+1)} \sum_{i=1}^q n_i (\bar{r}_i - \frac{n+1}{2})^2 = \frac{12}{n(n+1)} \sum_{i=1}^q n_i \bar{r}_i - 3(n+1)$$

8.5 כיצד נדע איזו קבוצה שונה?

נניח ומצאנו במבחן אנווה, כי קבוצה אחת שונה מהשאר.Cutת נרצה לדעת: איזו קבוצה (או איזה זוג) שונות משאר הקבוצות שגרמה לכך שנדחה את השערת האפס.

בעברית: השוואות אנליטיות. באנגלית - Post hoc tests .
 מה נוכל לעשות? וכל לבצע בדיקת השערת בין כל זוג. כמובן, נعبر על כל הזוגות האפשריים. נזכר - כי אנחנוCutcut עשויים בדיקות מרובות - אבל עליינו לתקן את הטעות $P - value$!
 מדוע שלא פשוט נשווה בין הממוצע של כל הזוגות של כל הקבוצות? נעלם את הסיכון לשגיאה מסוג ראשון, אז למה שלא נתקן? שכן זה יוצר בעיות אחרות.
 ישנה שיטה טובה יותר: Post hoc test דואג למנעו שגיאות שכאל. ישנים סוגים מבחנים ובים תחת השם Post hoc tests - Post hoc tests של בונופרי הוא אחד מהם. מה שונה בניהם הוא בהנחה היסוד שלן על הקלט והנתונים.

- נדון במבחן שכזה בשם HSD:
 א. נשווה כל זוג של ממוצעים לדוגמה $\mu_j - \mu_i$ לכל j, i .
 ב. הנחות יסוד:

1. הנתונים בלתי תלויים זה בזאת
2. הנתונים בכל קבוצה מותפלגים נורמלית
3. השונות בין הקבוצות דומה.

נבחין, כי אלו דרישות שדרשנו גם באנובה ולכן זה בסדר עבורנו כי הגענו לכך אחרי שעשינו אנווב.

המבחן שיקרא גם המבחן של Tukey מגדיר ערך: $q_s = \frac{|\mu_A - \mu_B|}{SE}$ - הערך המוחלט של המרחקים בין הממוצעים, מוגרם ב- $\frac{\sigma}{\sqrt{n}}$.

מהי התפוגות q ? דוגמאות n דוגמאות m אוכלוסיות בעלות התפוגות זהה (σ, μ, N). נסמן x_{min} כממוצע הקטן ביותר של אחת מן הקבוצות ואת x_{max} כממוצע הגדל ביותר של אחת מן הקבוצות, s^2 סטיית התקן של כל הקבוצות. אז הוא מגדיר $q = q_s = \frac{x_{max} - x_{min}}{s}$ שמתפלג בתפוגות q . הרעיון הוא להסתכל על כל זוגות הדוגמאות שיש בעולם "ולבדוק היכן זוג הדוגמאות הנ"ל נופל בתפוגות".

8.5.1 הפרש הממוצעים המינימלי שהוא מובהק סטטיסטית

ראינו כי $HSD = |\mu_A - \mu_B| = q_s SE = q \sqrt{\frac{s^2}{n}}$ ולכן, $HSD = \frac{|\mu_A - \mu_B|}{\sqrt{\frac{s^2}{n}}}$ והוא $\sqrt{\frac{s^2}{n}}$ באשר s הוא האמד שילנו לשונות האוכלוסיה). HSD הוא ההבדל המינימלי בין זוג ממוצעים שייחסו מובהק סטטיסטית (באוטו סוף בו חושב q) - ככלمر HSD יהיה הסף שבו אנחנו נגד - אם ההפרש ביןיהם יותר מאשר יש הבדל ביןיהם שגורם לפער באנווב.

8.5.2 קירוב של q

ישנה טבלה של q , ניתן להוציא את q ולהכפיל בסטיית התקן ולקבל את HSD . אם כן, ישנו קירוב של q . עבור 30 אנשים ומעלה (30 ומעלה דרגות חופש) מתקיים כי q מתפלג בקירוב כ- $\sqrt{2}z \approx \sqrt{2}z$ ו- z הוא סigma. ולכן נלק' לטבלת Z . וא'

$$HSD = z \sqrt{\frac{2 \times MSE}{n}}$$

באשר n הוא מס' הדוגמאות בכל קבוצה.

8.5.3 סיכום HSD

- א. חשב $ANNOVA$ - נרשום לפניו את הפיזור בין הקבוצות ואת מס' דרגות החופש.
- ב. חשב את הערך הקריטי של q . - על סמך מס' הקבוצות, מס' דרגות החופש באשר N הוא גודל המדגם ו- k מס' הקבוצות מס' דרגות החופש יהיה $k - N$ ורמות המובייקות הדרושים.
- ג. חשב את ההפרש המינימלי שייחס משמעותי סטטיסטית לפי $q \sqrt{\frac{SE^2}{n}}$
- ד. נשווה כל זוג הפרשים ונבדוק אם הוא מעיל הסף או שלא.

דוגמא:

נניח:
 4 קבוצות
 20 דגימות, 5 בכל קבוצה
 $MSE=10$

ברמת מובהקות של $df=20-4=16$, $k=4$, $N=20$, וLEN=0.05

לפי הטבלה המתאימה: $q=3.63$

$$HSD = 3.63 \sqrt{\frac{10}{5}} = 5.13$$

לכן, כל הפרש בין ממוצעים שגדול מ-5.13 הוא משמעותי סטטיסטי

8.6 אנוּבָה בְּשִׁתְיִ מְשֻׁתְּנִים

אנוּבָה מדברים כתע על מצב של: גברים מול נשים. האם יש הבדל משמעותית סטטיסטי בגובה של גברים לעומת נשים וגם לעומת גזע. כתע: כל קבוצה היא לפני 2 פרמטרים: גזע, ומין.
 דוגמה נוספת: השוואת גיל עצמו כי זה משתנה רציף).

ווכל לרשותם:

$$x_{j,k} = \mu + \alpha_j + \beta_k + \Delta_{j,k}$$

כל נקודה במדגם מוגדרת ע"י הממוצע הכללי של האוכלוסייה μ , ערך α_j (ערך של אותה קבוצה בפקטור הראשון - שינוי כתוצאה מקבוצה ראשונה), ערך β_k (ערך של אותה קבוצה בפקטור השני - שינוי כתוצאה מקבוצה שנייה) וערך $\Delta_{j,k}$ - שונות סביב הממוצעים.

כעת, נניח כי $\Delta_{j,k} \sim N(0, \sigma^2)$ (הפייזור סביב הנקודות הוא אויר). בה"כ $\sum_j \alpha_j = 0$ ($\sum_k \beta_k = 0$) (מדווע ניתן להניח זאת? תמיד אפשר להוריד את הערך שימושו להם ולהכניסו לממוצע הכללי). הנחות היסוד הן 2 כתע:

$$H_0^1 : \alpha_j = 0, H_0^2 : \beta_k = 0$$

כעת השערת האפס היא שלפי שני הפקטורים לא יהיה שינוי בממוצעים. השערת האפס כאן היא מוחמירה. וכך מספיק שהשערה אחת תופר בצדדי שנדחה את השערת האפס בכללותה.

כעת, מדובר בבדיקה באוטו פיתוח שעשינו בעבר אנוּבָה. נסתכל על הפיזור בתוך הקבוצות מול הפיזור של הקבוצות אחת יחסית לשנייה.

מקור השונות	hypothesis	degrees of freedom	סכום הריבועים	מדד סטטיסטי
בין קבוצות 1	$SS_{Between1} / SS_{Residual}$	a-1	$SS_{Between1}$	
בין קבוצות 2	$SS_{Between2} / SS_{Residual}$	b-1	$SS_{Between2}$	
בתוך הקבוצות		(a-1)(b-1)	$SS_{Residual}$	
סה"כ			SS_{Total}	

כעת, יש לנו שני מדדים סטטיסטיים ונעבור בדיקת כמו קודם - נבדוק את שתי השערות האפס, אם שתיהן לא יתנו לדחיה - סה"כ השערת האפס כולה תתקבל. (זה כמובן ניתן להכללה עבור יותר מ-2 משתנים וכן הלאה).

הערה. יתכן שיש תלות גם בין המוצא ל민. למשל: יש קשר בין גברים לאסיאתים. لكن יהיה חישוב שונה - הדוגמה מופיעה במצבת של הרצתה 10 בעמוד 46. הערה חשובה: אנו בהוא מבחן חד צדי ימני.

הרעיון הוא שניתן לחשב את האנובה שלנו לפי שלושה דרכי שביהם העולם יכול להתנהג. (שכן מתוכנה סטטיסטיית ומחישוב ידני התוצאות יכולות להיות שונות. מדוע? בדיק בಗל הדין זהה). זה נקרא *sum of squares*. הרעיון הוא - את מודל האנובה ניתן לחשב לפי כמה דרכי שביהם העולם יכול להתנהג: 1, 2, 3.

סוג 1: יש למשתנים שלנו סדר מסוים. למשל: החלוקה הכחיתו היא לפי הגז, ורק אחר כך לאחר מכן. או לחלופין. במקרה זה אנחנו מחשבים את SS לפי הסדר בו המשתנים מופיעים במודל. בסוג זה - מחשבים את *SOS* לפי המשטנה הראשון, ואת המשטנה הבא מחשבים מהשארית של מה שנשאר לאחר הורדת הראשון. משתמשים כאשר יש סדר שרצוי להתחשב בו.

סוג 2: לא משנה הסדר המופיעים, כל אחד יוחשב בפני עצמו. מחשבים את *SOS* בנפרד. משתמשים אם אין תלות.

סוג 3: אין תלות, מחשבים כל SS בפני עצמו + אינטראקציות ביניהם. ("בין הקבוצות"). משתמשים רק שיש נתונים תלויים או לא מאוזנים.

סוג	תלוי בסדר המשטנים	1	2	3
	כ. נחגג את ה- SS לא כל SS מחושב בפני עצמו. למשתנים פל סדר שבו מה פועעים מבודל	כ. נחגג את ה- SS לא כל SS מחושב בפני עצמו. אוטו-אינטראקציות	כ. נחגג את ה- SS לא כל SS מחושב בפני עצמו. אוטו-אינטראקציות	כ. נחגג את ה- SS לא כל SS מחושב בפני עצמו. אוטו-אינטראקציות
פצי' למשתנים האחרים	ר.ק. ממשתנים הקווומיים בסדר המודול	ר.ק. ממשתנים הקווומיים בסדר המודול	ר.ק. ממשתנים הקווומיים בסדר המודול	ר.ק. ממשתנים הקווומיים בסדר המודול
מתי משתמשים?	משנים לובי תליין, תלוות מושפעית או תנוונים אל אנדומים	משנים לובי תליין, תלוות	משנים לובי תליין, תלוות	משנים לובי תליין, תלוות
מתי לא להשתמש?	יש מתאמ ב- גורמים	יש מתאמ ב- גורמים	יש מתאמ ב- גורמים	יש מתאמ ב- גורמים

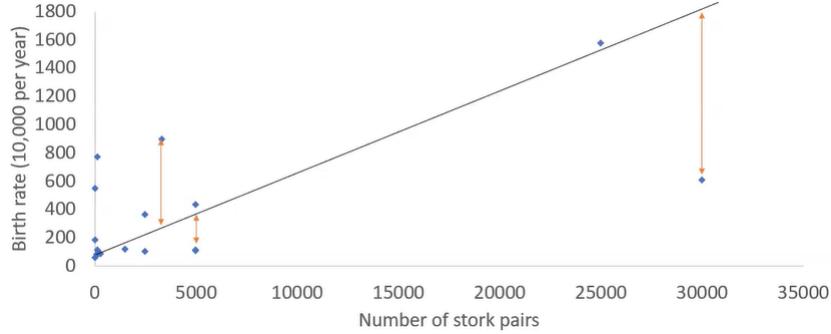
מה עושים אם הנתונים לא מפולגים נורמלית? ככלומר - אנחנו ביוטר ממשטנה אחד, האם יש מבוזן דמי קרטוסקל? יש, הוא לא מוגש מוצלח. לכן: לא נלמד אותה. ולכן מסקנה: אם הנתונים לא מפולגים נורמלית - אין לנו מבחן טוב עבור יותר ממשטנה אחד.

9 הרצתה 10: רגסיה לינארית

מדובר במשפחחה אחרונה של בדיקת אחרונות שנלמד, כעת אנחנו לומדים למדוד כיצד לבדוק השערות במשתנים רציפים. עד כה: למדנו על בדיקת השערות במשטנה אחד ובקבוצות קטגוריאליות, לאחר מכן הרחכנו באנובה על בדיקת השערות עם משתנה אחד ויתר מ-2 קבוצות (ואז גם הרחכנו למשתנים...). כעת נדבר על בדיקת השערות במשתנים רציפים.

נתבונן בדוגמה ראשונה. השערה: חסידות מביאות ילדים לעולם. הרי, באירופה מגיעות מאפריקה בקרבת האביב הרבה חסידות - ובאביב גם נולדים ילדים. אז, ברור שחסידות מביאות ילדים לעולם (וזו השערה שקשה לבדוק). מה אפשר לבדוק במקרים? האם ישנו קשר בין חסידות לילדים? ככלומר - האם ככל שיש יותר חסידות בעולם, ישנו יותר ילדים בעולם? איך ניתן לבדוק השערת שכזו? ראשית, כמובן שנאוסף נתונים. נתבונן במס' החסידות במדינה מסוימת (משטנה רציף עד כדי עיגול) ובמס' התינוקות באותו מדינה (גם משטנה רציף). ההיפותזה

שלנו - יש בינהם קשר. יש בינהם תלות (נניח לינארית). נ取 את הנקודות, ונסמן בגרף. נבחן כי אנחנו מעוניינים להעביר קו יקן מגמה.
איזה קו נעביר? **קו שסכום הריבועים ממנה יהיה מינימלי** (כלומר, המרחקים בערך מוחלט (בריבוע מן הסתם)()) מן הנקו. [ישנם קритריונים ערכיים שניית להחלטת, אנחנו נחליט על קритריון זה].



אנחנו כמובן, נרצה לחשב את הקו הזה. נפרמל:

9.1 הגדרה פורמלית - מציאת קו המגמה

נניח כי נתון לנו מוגדים ובו זוגות של נקודות (x_i, y_i) שאוון אנחנו רוצחים לקרב על ידי ישר. X מכוונים (הציר האופקי): משתנים בלתי תלויים, משתנים מסבירים או מאפיינים. Y מכוונים (הציר האנגץ): משתה תליי, targets.

משוואת הישר הינה $y = wx + b$ (w באשר לעיתים נסמן ϵ במסמל רעש) הערך על הישר בנקודה x_0 הוא (x_0, \hat{y}_0) מכאן, שהמරחק הריבועי של נקודה על הישר מהנקודה המותאמת במדגם הינו: $\epsilon_i = (y_i - \hat{y}_i)^2$ כיצד מוצאים את משוואת הישר לפי הクリיטריון הנ"ל?

$$E = \min\{\epsilon\} = \min\left\{\sum_i \epsilon_i\right\} = \min\left\{\sum_i (y_i - \hat{y}_i)^2\right\} = \min_{w,b}\left\{\sum_i (y_i - w \times x_i - b)^2\right\}$$

כיצד מוצאים מינימום של פונקציה? נגזר לפי b , w ונשווה לאפס:

$$\frac{dE}{dw} = 0 = \sum_i -2x_i(y_i - wx_i - b) = -2 \sum_i (y_i - wx_i - b)x_i$$

$$\frac{dE}{db} = 0 = -2 \sum_i (y_i - wx_i - b)$$

$$\sum y_i x_i = w \sum_i x_i^2 + b \sum x_i \quad \text{Mahmushoava haRishona Nekbel Ci}$$

$$b = \frac{1}{n} \sum y_i - \frac{1}{n} w \sum x_i \implies b = \bar{y} - w \bar{x} \quad \text{Mahmushoava haShniya Nekbel Ci}$$

נזכור למשוואת הרישונה,

$$\sum y_i x_i = w \sum_i x_i^2 + b \sum x_i$$

$$\sum y_i x_i = w \sum_i x_i^2 + (\bar{y} - w\bar{x}) \sum x_i$$

$$\sum y_i x_i = w \sum_i x_i^2 + \bar{y} \sum x_i - w\bar{x} \sum x_i$$

$$w(\sum_i x_i^2 - \bar{x} \sum_i x_i) = \sum y_i x_i - \bar{y} \sum x_i$$

ונקבל כי:

$$w = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum_i x_i^2 - n\bar{x}^2}$$

$$b = \frac{\bar{y} \sum x_i^2 - \bar{x} \sum x_i y_i}{\sum x_i^2 - \frac{1}{n}(\sum_i x_i)^2}$$

כעת, אנחנו יודעים בהינתן נתונים למצוא את w , b עבור קו המגמה שմבצע מינימום לסכום מרחק הריבועים.
כעת, נרצה לראות שיטות נוספות למציאת b, w . שיטות אלו מביאות אותה תוצאה, אך ימדו אותנו עוד.

9.2 שיטה שנייה (השיטה הסתברותית)

ונכל להסתכל על הבעה בראייה הסתברותית. קלומר:

$$\min \{E(Y - g(x))^2\}$$

משמעות: אם X ו- Y הם משתנים גאוסיאניים בעלי התפלגות משותפת $Variance$ $f(x, y)$ סופי, פתרון ריבועים מינימליים הוא קו רגסיה הנתון ע"י:

$$\frac{y - \mu_y}{\sigma_y} = \rho \frac{x - \mu_x}{\sigma_x}$$

$$\begin{aligned} \text{באשר } \rho &= \frac{\sigma_{xy}}{\sigma_x \sigma_y} \\ \text{ובמילים אחרות: } w &= \frac{\sigma_{xy}}{\sigma_x^2} \end{aligned}$$

9.3 שיטה שלישית (אלגברה לינארית)

בשיטת זו נשתמש כאשר יהיה יותר משתנה אחד. באלגברה לינארית זה הולך להיות הרבה יותר פשוט.

מה השיטה אומרת? נתבונן ראשית במטריצה בגודל $2 \times n$ [באופן כללי, אם נסמן את מס' המשתנים ב- t] מטריצה בגודל $t+1 \times n$. נשים במודה השניה ערך קבוע כלשהו, למשל: 1:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_{n-1} \\ y_n \end{pmatrix}$$

וכן נתבונן בוקטור עמודה בגודל $1 \times n$ בשם b .

X	b
x_1	1
x_2	1
..	..
x_n	1

לנו $1 - m$ משתנים איזי X היא מטריצה בגודל $m \times n$ כך:

$$\begin{pmatrix} x_1 & z_1 & t_1 & .. & 1 \\ x_2 & z_2 & t_2 & .. & 1 \\ .. & .. & .. & .. & 1 \\ x_{n-1} & z_{n-1} & t_{n-1} & .. & .. \\ x_n & z_n & t_n & .. & 1 \end{pmatrix}$$

נרצה למצוא וקטור W כך $W = XW$. באופן כללי: Y הוא וקטור עמודה באורך n , X היא מטריצה בגודל $m \times n$ ו- W הוא וקטור עמודה בגודל m (כמס' המשתנים הבלתי תלויים + 1). הרעיון יהיה: בכפל מטריצות אחד - לסייע את הכל: ולא להכנס להרבה חישובים. בשביל למצוא את W , נדרש למצער את :

$$E = (Y - XW)^T(Y - XW)$$

אבל - זו מטריצה! כיצד אנחנו גוזרים מטריצה? קלומר מהו $\frac{\partial E}{\partial W}$?

נגזרת של מטריצה

נדיר:

$$\frac{\partial y}{\partial x} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & .. & .. & .. & \frac{\partial y_1}{\partial x_n} \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ \frac{\partial y_m}{\partial x_1} & .. & .. & .. & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

כלומר: אם רוצים לגזר את המטריצה Y לפי x , כל תא נגורז בהתאם למשתנה x המתאים בו.

דוגמא: אם A היא מטריצה סימטרית, (למשל $X^T X$ תמיד סימטרית) איז עבור הצורה הריבועית $z^T A z$ הנגזרת לפי z הינה:

$$\frac{\partial}{\partial z}(z^T A z) = 2A z$$

מדוין? $\frac{\vartheta(z^T A z)}{\vartheta z_k} = \sum_{i=1}^n z_i A_{ik} + \sum_{j=1}^n A_{kj} z_j$. וכך, $z^T A z = \sum_{i=1}^n \sum_{j=1}^n z_i A_{ij} z_j$. $2 \sum_{i=1}^n z_i A_{ij} = 2 A z$

cut, נחזר למשמעות ערך E שלו:

$$E = (Y - XW)^T (Y - XW) = (Y^T - W^T X^T)(Y - XW) = Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW$$

$$Y^T Y - 2W^T X^T Y + W^T X^T XW$$

שכן המעבר מהשורה הראשונה לשניה, התאפשר בזכות העובדה ש $Y^T XW = W^T X^T Y$ כי מדובר בסקלר ולכן שמספריים טרנספוז זה לא משתנה. נגזר את הפונקציה שלנו, ונשווה לאפס:

$$\frac{dE}{dW} = -2X^T T + 2X^T XW = 0$$

$$X^T XW = X^T Y$$

(הגירה הטעינה לפי הכלל קודם, שהרי $X^T X$ סימטרית ולכן הביטוי $W^T X^T XW$ יגור להיות $2X^T XW$ לפי הכלל הזה). $\frac{\vartheta}{\vartheta z_k}(z^T A z) = 2A z$ קלומר המינימום של הפונקציה הינו: $X^T XW = X^T Y$. כיצד נבודד את W ? נכפול בשני הצדדים ב $(X^T X)^{-1}$

$$W = (X^T X)^{-1} X^T Y$$

למשמעות זו יש שם: *Pseudo inverse*. נבחן כי המימד של $X^T X$ הוא $m \times m$ - ולכן השם: די קטן. [נעיר כי אי אפשר לפתחו לפי חוקי מטריצות, שכן X אינה מטריצה ריבועית אך $X^T X$ כן כן].

$$A^{-1} = \frac{1}{|A|} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \text{ איז}$$

מכאן, המשוואה הנ"ל נותנת לנו נוסחה סגורה לחישוב קו הרגרסיה W . מדובר כמובן להשתמש באלגברה ליניארית בשbill לפטור את מודל הרגרסיה (ולא במשוואת מהמודול הראשוני?) שכן: בכל מצב שיש יותר משתנה בלתי תלוי אחד, המשוואות רק מתחילה להסתבך. וכמובן (!) להשתמש בשיטה זו.

לבסוף, לאחר שהיחסנו את W קיבל את קו הרגרסיה. זה, יש קשרلينאר? נניח וקיים קו את הקו $y = 0.0287 + 233.76 x$ - זו התאמת טובה? כיצד נדע לקבוע מהו טוב ומה לא טוב?

9.4 האם הקשר לינארי?

קיבלנו את קו הרגרסיה. מה הלאה?
זכור במקדם המתאים של פירסון - שקיימים $1 \leq r \leq -1$. אם $-1 \leq r = 1$, אם $0 > r$ ישנו קשר לינארי חיובי בין המשתנים ואם $0 < r$ ישנו קשר לינארי שלילי.
נבחן כי לפי הנוסחה:

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

נתבונן ב- r^2 :

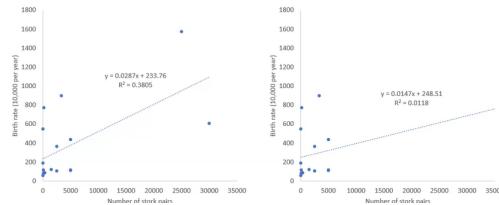
$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{w^2 \sigma_x^2}{\sigma_y^2} = \frac{w^2 E(x - \bar{x})^2}{\sigma_y^2} = \frac{E((b + wx - \bar{y})^2)}{\sigma_y^2} = \frac{E(\hat{y} - \bar{y})^2}{E(y - \bar{y})^2} = \frac{SS_{REG}}{SS_{TOT}}$$

שכן המעברים לפי מה שראינו Ci $w = \frac{\sigma_{xy}}{\sigma_x^2}$
כעת: SS_{REG} הוא סכום הריבועים שמשבירים כמה שונות נשאר בנתונים לאחר שהעבכנו את קו הרגרסיה. במכנה, SS_{TOT} הוא סכום הריבועים על הנתונים המקוריים (כלומר כמו y סטistics מה ממוצע שלהם). לכן: r^2 הוא אחוז $variance$ שמוסבר ע"י הרגרסיה.
אם קיבלנו Ci $r^2 = 0$ אז \hat{y} תמיד שווה לע' לא משנה מה נעשה. זה לא טוב! לעומת זאת: אם $r^2 = 1$ זה אומר Ci $SS_{REG} = SS_{TOT}$ שכן כל פעם שנלך לאיזושהי נקודה במדגם, $\hat{y} = y$ בRibou שווה ממש למה שיש במדגם המקורי. וכך: הקירוב ממש מצוין.
משמעות: אם $0 < r^2$ הרגרסיה לא מסבירה כלום. אם $1 = r^2$ הרגרסיה מסבירה הכל באופן מושלם.

כעת, אם נחזור לדוגמה הקודמת: $y = 0.0287x + 233.76$. מה זה אומר? זה אומר 38% מהשונות מוסברים ע"י הרגרסיה. האם זה הרבה? זה תלוי - בכמה אנחנו מעוניינים להסביר. יתכן ואלעד יגיד ש-38% מספיק לו בשbill להגיד: יש קשר בין החסידות לילדים. מישחו אחר יגיד לו - לא הסברת 62% אחרים מהשונות.

חשוב לשים לב' הרגרסיה הגיע מודל נתונים חריגים. אם יש לנו שני נתונים חריגים מאוד מהשאר - הם מאוד מאד ישפיעו לנו על r^2 .

בדוגמה הבאה: לאחר הסרת 2 הנתונים 38% מהנתונים ירדנו להסביר על 1% בלבד:



9.5 חישוב מובוקות סטטיסטית

מה היינו עושים אם היה לנו יותר ממשתנה אחד? במקרה זה אי אפשר לציר גורם עם שיפוע Ci יsono יותר ממשתנה אחד. לא נכון לומר שישנו קורלציה. מה נעשה? נרצה למצוא דרך לעשרות בדיקות השערות - ולבדוק האם השיפוע של כל משתנה ב"ת מכלל המשתנים הב"ת שיש לנו שונה מפות בצורה משמעותית או שלא?
נכחות השערות:

$$H_0 : w_i = 0, H_1 : w_i \neq 0 (w_i <> 0)$$

כלומר - השערת האפס היא שהשיפוע שמצאנו של המשתנה ה i שווה לאפס. והאלטרנטיבית: הוא שונה מאפס.

כיצד נעשה זאת? נסתכל על הסטטיסטיק $\frac{\hat{w}_i - 0}{SE(w_i)}$.
 הוריאנס של השארית ε מוחשב ע"י $\frac{\|y - XW\|^2}{n-k} = \frac{RSS}{n-k}$ (כיוון שהראש ב"ת בכל המדידות ניתן כתוב כי $y \sim N(0, \sigma^2 I)$ ע"י המרחק הריבועי בין y לשערות שלו בכל אחת מהמדידות שלנו). כלומר,

$$SW(E) = \sqrt{diag(\sigma^2(X^T X)^{-1})}$$

לכן: $t = \frac{\hat{w}_i}{SE(w_i)}$ עם $n - k - 1$ דרגות חופש. באשר n הוא מס' הנתונים ו- k מס' המשתנים הבלתי תלויים (כולל הקבוע). ככל מרובה, מסתכלים על ערך זה בהתקפלות t ועושים מבחן t עם $1 - \alpha$ דרגות חופש, ורגיל: ברמת מובהקות מקבלים או שודחים.

הרצאה 12: AB TESTING

הרצאה 13: חזרה למבחן