

שיטות סטטיסטיות - סיכום הרצאות לבחן

20 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 | הרצאה 1: מבוא לקורס | 1 |
| 3 | שיטות מחקר: | 1.1 |
| 4 | מעגל החיים של ניסוי: | 1.2 |
| 4 | הרצאה 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 |
| 5 | לשם מה אנחנו אוספים את המידע? | 2.1.3 |
| 5 | שיטות דוגמה הסתברותיות | 2.1.4 |
| 5 | שיטות דוגמה לא הסתברותיות | 2.1.5 |
| 5 | משתנים | 2.2 |
| 5 | סוגי משתנים | 2.2.1 |
| 6 | סיווג משתנים - סולמות מדידה | 2.2.2 |
| 6 | מדדים סטטיסטיים | 2.2.3 |
| 6 | תיאור והצגה | 2.3 |
| 9 | הרצאה 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 |
| 13 | שונות וסטיתת תקן | 3.4 |
| 14 | מוצע משוקלל ושונות מצורפת | 3.5 |
| 14 | מדדי קשר בין מספר משתנים | 3.6 |
| 15 | הרצאה 4: הסקה סטטיסטית | 4 |
| 15 | מבוא | 4.1 |
| 15 | הסקה סטטיסטית | 4.2 |
| 16 | מושגים בסיסיים | 4.3 |
| 16 | התפלגות דוגמה | 4.4 |

| | | |
|----|---|--------|
| 18 | הרצאה 5 : אמידה סטטיסטית נקודתית | 5 |
| 19 | אמידה סטטיסטית | 5.1 |
| 19 | בעיית האמידה | 5.2 |
| 20 | תכונות של אמידים | 5.3 |
| 22 | שיטות אמידה | 5.4 |
| 22 | שיטת המומנטים | 5.4.1 |
| 24 | שיטת הניראות המרבית | 5.4.2 |
| 25 | הרצאה 6 : אמידה סטטיסטית של מרוחבי בטחון | 6 |
| 26 | רוחם סמך של ממוצע המדגם | 6.1 |
| 26 | רוחם סמך - הגדרה פורמלית | 6.2 |
| 27 | רוחם סמך לממוצע עבור התפלגות נורמלית כאשר השונות ידועה כאמור לתוכלה | 6.3 |
| 28 | רוחם סמך לממוצע כאמור לתוכלה כאשר השונות אינה ידועה | 6.4 |
| 30 | הרצאה 8 + 7: הסקט סטטיסטית - בדיקת השערות | 7 |
| 30 | השערת האפס וההשערה האלטרנטטיבית | 7.1 |
| 31 | סוגי השגיאות | 7.2 |
| 31 | חישוב הסבירות של השערת האפס | 7.3 |
| 32 | שלבים להחלטה אם לקבל או לדחות את השערת האפס | 7.4 |
| 32 | תיקון ל מבחנים מרובים | 7.5 |
| 33 | חישוב גודל המדגם הדרושים | 7.6 |
| 33 | מספר מוגדים | 7.7 |
| 34 | גודל האפקט (d של כהן) | 7.8 |
| 34 | סיכום חשוב | 7.9 |
| 34 | מבחני השערות במוגדים גדולים | 7.10 |
| 35 | מבחני השערות | 7.11 |
| 35 | דוגמה ראשונה: ממוצעים | 7.11.1 |
| 36 | דוגמה שנייה: הצלחות במוגדים | 7.11.2 |
| 36 | דוגמה שלישיית: הפרשים בין ממוצעים | 7.11.3 |
| 38 | מבחנים של זוגות | 7.11.4 |
| 39 | מבחנים אפרמטריים: מבחן χ^2 | 7.11.5 |
| 39 | מדידת הפרשים א-פרמטרית במוגדים בלתי תלויים: מבחן <i>Mann Whitney U TEST</i> | 7.11.6 |
| 42 | מדידת הפרשים א-פרמטרית בדגם מזוג (מבחן <i>Wilcoxon Singed-Rank</i>) | 7.11.7 |
| 43 | סיכום מבחני השערות | 7.12 |
| 44 | הרצאה 9: <i>(Analysis of Variance) Anova</i> | 8 |
| 45 | הנחות יסוד לבחן אונובה (בסיסי) | 8.1 |
| 45 | הרעיוון המרכאי של מבחן אונובה | 8.2 |
| 45 | הגדרה פורמלית | 8.3 |
| 49 | כיצד מתגברים על ההנחהות? | 8.4 |
| 49 | איך מתגברים על ההנחה שהנתונים מתפלגים נורמלית? | 8.4.1 |
| 49 | כיצד נדע איזו קבוצה שונה? | 8.5 |
| 50 | הפרש הממוצעים המינימלי שהוא כובח סטטיסטית | 8.5.1 |
| 50 | קירוב של q | 8.5.2 |
| 50 | סיכום <i>HSD</i> | 8.5.3 |
| 51 | אונובה בשתי משתנים | 8.6 |

| | | |
|----|---|--------|
| 52 | הרצאה 10: רגרסיה לינארית | 9 |
| 53 | הגדירה פורמלית - מציאת קו המגמה | 9.1 |
| 54 | שיטת שנייה (שיטת ההסתברותית) | 9.2 |
| 55 | שיטת שלישיית (אלגברה לינארית) | 9.3 |
| 57 | האם הקשר לינארי? | 9.4 |
| 57 | חישוב מובהקות סטטיסטיות | 9.5 |
| 58 | הרצאה 12: רגרסיה מותקמת | 10 |
| 58 | רגרסיה לא לינארית | 10.1 |
| 58 | רגרסיה של סדר (Rank) | 10.2 |
| 59 | תיקון R^2 לגודל המדגם | 10.3 |
| 59 | <i>Identifiability</i> | 10.4 |
| 59 | <i>Variance Inflation Factor</i> | 10.5 |
| 60 | <i>Ridge Regression</i> | 10.6 |
| 60 | מה עושים אם חלק מהנתונות היסוד של רגרסיה לא מתקיימות? | 10.7 |
| 61 | הרצאה 13 : <i>AB TESTING</i> | 11 |
| 61 | הגדירה | 11.1 |
| 62 | סוגי מדדים (OEC) | 11.1.1 |
| 62 | קריטריונים למדדים | 11.1.2 |
| 63 | השפעת הניסוי על המשתתפים | 11.1.3 |
| 63 | הרצת ניסויים במקביל | 11.1.4 |
| 63 | טעויות נפוצות בניסוי <i>AB</i> | 11.1.5 |

1 הרצאה 1: מבוא לקורס

סטטיסטיקה: תחום ידע שנוגע לאייסוף, עיבוד נתונים והסקת מסקנות מנתונים כמותיים. מחלקים את הסטטיסטיקה לשני תחומי דעת: תאוריית, והיסקית.

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

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

AMPLEDAH STATISTIKA: אלו שיטות מתחממות שמאפשרות לנו רוח מתחום נתוני המדגם אומדן ערך של משתנה עבור אוכלוסייה. הבסיס למידה חישובית.

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

AMPURI = 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 תמורה, ורעות היא בת אדם, גם רעות היא בת תמורה).

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

הגישה המדעית = הגישה האמפירית + הגישה הרצינולית.

מטרת הגישה המדעית: להבין עבר, לנבأ עתיד ובעיקר לנתח תאוריות.
תאוריה מדעית: מערכת מונחים, הגדדות וטענות. התאוריה כוללת מערכת של טענות על קשרים בין מונחים.

ניסוח בעית מחקר: בעית מחקר היא בעיה שניתן 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 בבחינה אמפירית (דרך למדידת משתנים) השערת מחקר - מומקדת וספציפית, משקפת את ציפיות החוקר וכן יש את קריטריון ההפרכה: השערה שיש דרך אמפירית להפריך אותה - מערך ניסוי.

1.2 מעגל החיים של ניסוי:

איסוף מידע (הרצאה 2) ← → תיאור והציגה (הרצאה 3) ← → אומדן פרמטרים (הרצאה 4 – 6) ← → בדיקת השערות (הרצאה 7 – 9) ← → ניסוח השערה ← → בחור על עצמן
אומדן מידע + תיאור והציגה = סטטיסטיקה תאורית.
אומדן פרמטרים + בדיקת השערות = סטטיסטיקה היסקית.

2 הרצאה 2: סטטיסטיקה תאורית 1

2.1 איסוף מידע

2.1.1 מיי אוספים את המידע?

א. **אוכלוסייה** – אוסף של אנשים, דברים, האובייקטים אותם אנו רוצים לחזור.

ב. **מדגם** – תת קבוצה (מייצגת) של האוכלוסייה
1. אם קיימים קשיים במידידה של האוכלוסייה כולה (מסובכת, ארוכה, יקרה)
2. קשיים באיסוף המידע (הרבה מידע)
3. עצם המודידה פוגע בתכונה (כלומר, אם למשל אנחנו במבצע גפרורים ורוצים לבדוק את מס' הגפרורים התקיימים – בשביל לבדוק אם הוא תקין נצטרך להשתמש בו ולכן הפכנו אותו ללא תקין. אם נkeh את כל הגפרורים ונבדוק אותם נשאר ללא גפרורים תקינים: لكن אנחנו חiyבים לחתת מדגם).

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

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

2.1.2 מה אנחנו אוספים?

משתנה: תכונה הניתנת ל特派ית ומידה עבור כל אלמנט באוכלוסייה.

ערך: הערך שנמדד עבור אלמנט יחיד באוכלוסייה.

מידע: הערכים שנמדדו עבור כל האוכלוסייה.

2.1.3 שם מה אנחנו אוספים את המידע?

סטטיסטי: ערך המוחשב על סמך הדטא, כולל על סך כל הערכים שנמדדו. (ממוצע הדגימות).

פרמטר: מאפיין של האוכלוסייה. למשל, תוחלת ההתפלגות.

–♥– המשנה הוא תכונה, למשל אם נבעצ' מוגדים סכום הכספי הממוצע שסטודנט מוציא בשנה א', וקיבלו שהממוצע הוא \$ 178, אז הסטטיסטי הוא \bar{x} וכן המשנה הוא סכום הכספי הממוצע.

–♥– **יתכן סטטיסטים שונים:** למשל מינימום מקסימום, חיצון, שונות וכו'.

2.1.4 שיטות דגימה הסטברותיות

בשיטות דגימה הסטברותיות 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. **בקורס זה כאמור כי אין מודדים – נמדד לפי דגימה אקראית.**
2. דגימה בשכבות – חלוקת האוכלוסייה לשכבות זרות ומשילומות. דגימה רנדומית (לפי פורפරציה) מכל שכבה. דוגמה לדגימה בשכבות: סקרי בחירות. לוקחים שכבות אוכלוסייה – אם יודעים שישנם 27% מהאוכלוסייה בגילאים 40 – 50 – או דוגמים פורפראציונליות משכבות גיל זו.
3. דגימת אשכולות – חלוקת כל האוכלוסייה לקבוצות זרות ומשילומות. דגימה רנדומית של קבוצות והוספת כל הפרטים מכל קבוצה למוגם. למשל: ביצוע סקר מדד האושר. במקום למדוד אחד, אפשר למדוד בתים אב. אם בית אב יצא $5/5$ במדד האושר – כל האנשים בבית האב היל' ייחשבו $5/5$ במדד האושר. כולל – לוקחים את כולם.

2.1.5 שיטות דגימה לא הסטברותיות

ישנן שיטות דגימה שאינן הסטברותיות.

1. **דגימת נוחות** – "מן המוכן", הכל בבת אחת. כולל – מקבלים את הדגימה בבת אחת. למשל: משאל רחוב, מקבלים את התוצאות מיד. מה טוב בשיטה? מהיר. מה בעייתי? 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 ולבן זה טוב לו שהיא בוחרת את האוכלוסייה. מה לא טוב? לא מייצג סיכוי גבוה להטיה.

3. **דגימת בדור שלג** – "חבר מביא חבר". כולל – ניסויים שאים מגע, מקבל כסף על הניסוי ואומרים לו להביא חברים לניסוי שיבוא גם הוא "להרוויח כסף". יתרון: קל ומהיר, דגימת אוכלוסייה זהה. חסרון: לא מייצג, ישנה הטיה, מוגם של חלק ספציפי באוכלוסייה.

ישנן **騰場** לתקופות הניסוי: דגימה לא מייצגת / מוגה, דגימה "התנדבותית", דגימה קתינה מדי. למה להשתמש בשיטות דגימה לא הסטברותיות? פיזיולוג, מיעוט אקטואטי, תופעות מאוד נדירות.

בקורס זה נשתמש בשיטות דגימה הסטברותיות.

2.2 משתנים

2.2.1 סוגי משתנים

א. **קטגוריאי:** קבועות ערכיהם סופית. קטגוריה מדרגתית, דרגה בצבא, קבועות המידע = $\{XS, S, M, L, XL, XXL\}$.

ב. **מספריאי:** מס' הסטודנטים בקורס, מספר אסיטטים למשחק, גובה משקל וכו'.

המשנה הבדיקה: קבועות ערכיהם סופית וbeit מניפה.

המשנה הרציף: קבועות ערכיהם אינסופית, בין שני ערכים קיימים ערך. למשל – מרחק.

2.2.2 סיווג משתנים - סולמות מדידה

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

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

סולם רוחניים: עם יחס סדר, עם מרווחים קבועים. למשל: טמפרטורה - משתנה מספריאי. בסולם זה: יש משמעות למרווחים בין הערכים. למשל בטמפרטורה יש משמעות למרווחים בין הטמפרטורות השונות.

סולם מנתה: יחס סדר, מרווחים קבועים, נקודת אפס. למשל: גובה, משקל. - משתנה מספריאי. 0 מייצג העדר תכונה.

2.2.3 מדדים סטטיסטיים

סטטיסטי: ערך המוחושב על סמך התצפיות בפועל. למשל - ממוצע, חציון.
 פרמטר: תוכנה של האוכלוסייה המקורית. למשל - תוחלת בהתפלגות נורמלית, פורפרציה בהתפלגות בינומית.
 ♡- בחלק של "סטטיסטיקה תאורית", נשתמש בסטטיסטיים לתיאור הדאטה. בחלק "הסקה סטטיסטית"
 נשתמש בסטטיסטיים כאומדן לפרמטרים.

2.3 תיאור והציג

כיצד ניתן להציג את המידע שנאסף?

* צגנה 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 . בטבלה מטה סה"כ שכיחות שמסתכמת ל-20. שכיחות יחסית תהיה האחוז של הערך v

| ערך v | שכיחות $f(v)$ | שכיחות 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 ייחסית, רק כל ערך כובר את השכיחות של הערך הקודם:

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

4. משתנה מספרי בדיד: חלוקה למחלקות

ניתן לחלק את הערכים השונים למחולקות. למשל במקומות להציג 10, 9, ..., 1, להציג כ- 4 – 7, 8 – 10 במחולקות שונות. לשם כך צריך לדאוג שהמחלקות יהיו זרות, חלוקה מסוימת שαιיחודם הוא כל ערכי המדגם ושמירה על גבולות דמיוניים בין המחלקות.

5. משתנה מספרי רציף: חלוקה למחולקות

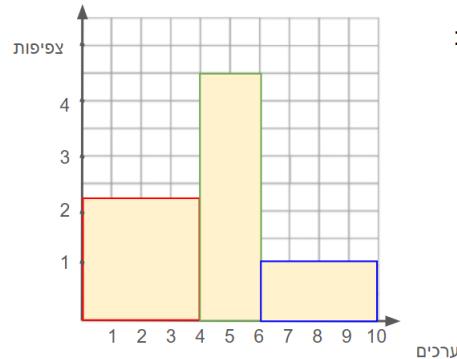
הגבול העליון של מחלקה אחת מתלכד עם הגבול התחתיו של זו אחרתיה. בהינתן מחלקה $[x_0, x_k]$ רוחב מחלקה: ההפרש בין גבול עליון אמיתי לגבול תחתיו אמיתי. יתקיים $I = x_k - x_0$
מחלקה פתוחה: רק גבול עליון או תחתיו

| מחלקה | שכיחות f | רווח מחלקה | מרכז טווח | כפיפות f=d |
|-------|----------|------------|-------------|------------|
| 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 |

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

כפיפות המחלקה מוגדרת להיות:

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



כיצד בונים היסטוגרמה?

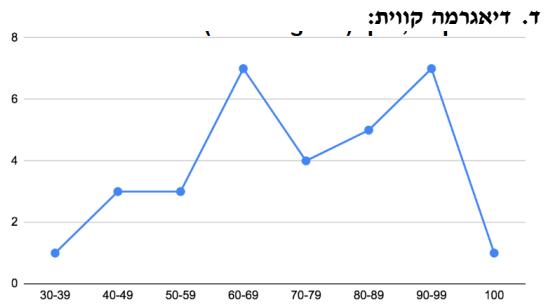
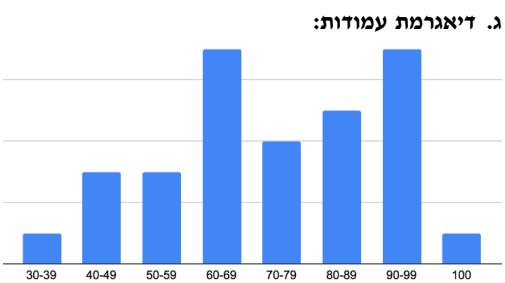
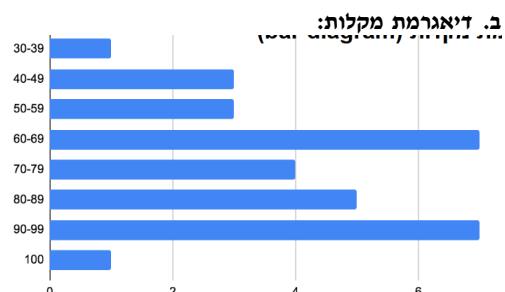
- מחליטים על מס' המחלקות שנרצה - k
 - מחשבים את הטווח של ההיסטוגרמה $b = \max - \min + 2$ (\max מוסיפים פלוס 2 רק באשר אנחנו ידעים את הערכים עצם ממש ולא היסטוגרמה.)
 - מחשבים רוחב כל מחלקה $a = \frac{b}{k}$
 - מחשבים גבולות מדוימים - $\min - a$
 - בחירת ייחידת הדיקוק u
 - чисוב גבולות אמיתיים $u - \min$
- ההיסטוגרמה נcona רק כאשר נתונים לנו כל הנתונים. אם נתון לנו טבלת שכיחויות אי אפשר לעשות זאת.

* תצוגה גרפית:

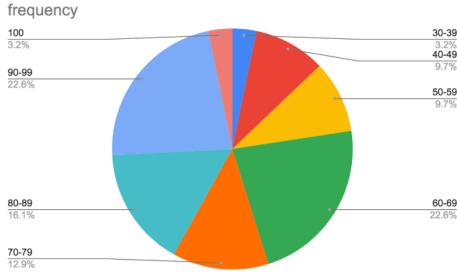
מיצגים נתונים באמצעות דיאגרמה.

א. דיאגרמת גבעול-עליה: דיאגרמה בה מפצלים את הערכים לעשרות ויחידות. חלוקת טווח הערכים לבעולים. וכן פירוט ערכי העלים.

| stem (מספר העשרות) | 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 6 |
| 10 | 0 |



ה. דיאגרמת עוגה:



מתי השתמש באיזו דיאגרמה?

עבור כל סולמות המדידה: טבלת שכיחיות, דיאגרמת עמודות, תרשיס עוגה.
עבור סולם רוחים או סולם מנה: היסטוגרמה, דיאגרמת גבעול עליה, דיאגרמת קופסה.

* מדדים סטטיסטיים:

שכיח: הערך עם השכיחות הגבוהה ביותר.

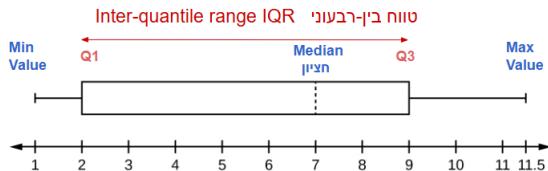
מרכז הטוחה: הממוצע בין התצפית הגבוהה ביותר והנמוכה ביותר.

חציון: 50% או פחות (אם לא זוגי) גבויים ממנו, 50% או פחות נמוכים ממנו.

ממוצע: סכום כל הערכים מחולק במספר התצפויות.

דיאגרמת קופסה:

בשביל לחישבה מסתכלים על ערך המינימום, המקסימום, החציון וכן Q_1 ו- Q_3 שייהי החציון של החציון הנמוך (מהמינימום עד החציון) ו- Q_2 שייהי החציון של החציון הגבוה (מהחציון אל המקסימום).



כיצד מציירים דיאגרמת קופסה?

א. מסדרים את הנתונים לפי סדר עולה

ב. מוצאים מינימום, מקסימום, חציון, רביעון ראשון ורביעון שני

ג. מצאים לפי הנתונים שמצאננו קודם – בין הרביעון הראשון לרביעון השלישי שלישית קופסה, בתוכה מסמנים את החציון.

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

הערה: למיציאת החציון – אם יש לנו מס' זוגי זה קל, האמצעי. אם יש מס' זוגי יש שני חציונים

- החציון בקורס יוגדר להיות ממוצע שני החציונים הנ"ל.

3 הרצתה 3: סטטיסטיקה תאורית 2

3.1 מדדים סטטיסטיים

סטטיטיסטי הסדר: יהיו X_1, \dots, X_n משתנים מקרים עבור אוכלוסייה או מדגמים. יהיו x_1, \dots, x_n הערכים שנמדדו עבורם בהתאם. נסדר את הערכים x_1, \dots, x_n בהתאם בסדר עולה. קיבל את סטטיטיסטי

הסדר:

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

שבייה: הערך עם השכיחות הגבוהה ביותר,

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

מרכז הטווח: הממוצע בין התצפויות הנמוכת 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 הערכאים חלקים מט' התצפויות

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

3.2 חישוב מדדי מיקום מרכז עבור מחלקות עם גבולות אמייניטיים

כאשר נתונה לנו טבלת שכיחויות וגבולות אמייניטיים, נסמן $f(v)$ כשכיחות של v , את השכיחות היחסית $r(v)$ ואת היחסית המצטברת $RF(v)$ כיחסית הממוצע \bar{x} :

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

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

כיצד נחשב את החציוון בטבלת מחלקות רגילה? נסתכל על השכיחות היחסית המצטברת $RF(v)$ ונחפש היכן אנחנו פחותים מהקלט, והחציוון יהיה שורה אחת אחריו. אם קיימן ערך עבורו $RF(v) = 0.5$ אז החציוון יהיה הממוצע של זה לפניו וזה לאחריו.

ומיצד נחשב חציוון עבור מחלקות עם גבולות אמייניטיים?
מחלקה m , גבולות $-L_1, L_0, L_1 - f$, שכיחות F ומצטברת

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

הרענון יהיה למצוא את האיבר אשר הכו מתחתיו מחלק את ההיסטוגרמה לשני חלקים שווים שטח. בשלב הראשון נצטרך להזות את המחלקה m בה החזון אמור להמצא, את זה נעשה כמו שעושים בטבלה רגילה. נסמן את הגבולות שלה ב- $L_0 - L_1$. וכן n מס' התכפיות.

באופן דומה, לחשב את הרבעונים: נזהה את המחלקה m בה נמצא הרביעון ונחשב \bar{x} נשים לב כי F הינה שכיחות מצטברת (לא יחסית!).

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

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

עבור מאון k :

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

אלפין 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 אמיטיים (גבול עליון של מחלקה קודמת זהה לנגבול תחתון של מחלקה נוכחית).

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

אם נבחר במדד מסוים, מהי פונקציית ההפסש שלו?

א. מס' השגיאות: כמה מהערכים אינם שווים למדד עצמו $|\{\bar{x}_i | x_i \neq \bar{x}\}|$:
כאשר נסתכל על פונקציה זו, השגיח ימזרע את מס' השגיאות. כמובן אם הפונקציית הפסד שמעניןית אוטו היא מס' השגיאות אי נשותמש בשיכחה.

ב. השגיאה המקסימלית: המרחק המקסימלי מהמדד עצמו $\max_i |x_i - \bar{x}|$

כאשר נסתכל על מדד זה, מרכז הטוחה מזעיר את השגיאה המקסימלית.

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

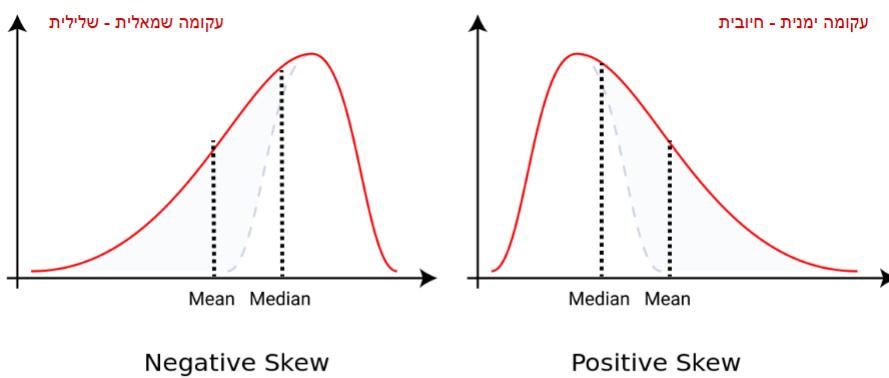
ד. סכום ריבועי השגיאות: מרחקים ריבועיים של כל הערכים מהמדד $\sum_i (x_i - \bar{x})^2$
המשמעות מפחית למינימום את סכום ריבועי השגיאות.

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

תכונות מדדים סטטיסטיים למיקום מרכזי \bar{x} :

| משמעות | שכיח | ממוצע טווח | חציון | ממוצע |
|-------------------|--------------|-----------------------|--------------|---------------------|
| פונקציית הפוד | מספר שגיאות | סכום השגיאות המוחלטות | סכום השגיאות | סכום ריבועי השגיאות |
| רגשות לערכי קיזון | רבה | רבה | אין | רבה |
| סולמות מדידה | רווחים ומעלה | רווחים ומעלה | שמע ומעלה | רווחים ומעלה |
| שימושיות בהסקה | רובה | רובה | בינונית | רובה |

נשים ל.ב. בעקבות פעמו סימטרית: הממוצע=חציון=שכיח.
 בעקבות פעמו אי סימטרית שמאלית (האנט לצד שמאל) : ממוצע > חציון > שכיח
 בijkומת פעמו אי סימטרית ימנית (האנט לצד ימין) : ממוצע < חציון < שכיח



3.3 מזרדי פיזור

- אחוֹ השגיאות: אחוֹ התצפויות השונות מהשכיח $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}| \neq \text{אחוֹ} \{i | x_i = \bar{x}\}$
- ב. גודל השגיאה המקסימלית: המרחק הגדול ביותר ממרכז הטווח $\max_i |x_i - \bar{x}|$
- ג. הטווח: המרחק בין ערכי קיזון
- ד. הטווח הבין רבוני: הטווח בו נמצאים 50% הערכים המרכזיים בתפלגות. (מה שאנוanno מציירים בדיאגרמת Box).
- ה. ממוצע הסטיות המוחלטות: ממוצע מרחקי התצפית מהחציון. $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$
- ו. ממוצע ריבועי הסטיות: ממוצע ריבועי מרחקי התצפית מהממוצע $\frac{1}{n} \sum_{i=1}^n |(x_i - \bar{x})|^2$. נשים ל.ב. כי הعلاה בריבוע מענישה יותר את הקצוות, וזה יותר טוב עבור המרץ ולכן זה בודק היפוטה את הקצוות.

תכונות:

| השגיאות | אחד | משמעות מקסימלית | טוה | טוה | טוה | משמעות בירכובני | משמעות סטיוית מוחלטת | משמעות סטיוית ריבועית | משמעות סטיוית | פונקציות הפסד |
|---------------------|---------------------|---------------------|---------------|---------------|---------------|-----------------|----------------------|-----------------------|---------------------|---------------------|
| שגיאות מקסימלית | מספר שגיאות | שגיאה מקסימלית | — | — | — | — | מוחלטות סטיוית | ריבועיות סטיוית | סכום סטיוית ריבועית | ריבועיות סטיוית |
| שליח | מרכז טוה | מרכז טוה | — | — | — | — | חץין | מוצע | מוצע | מדד מדן המרכז הבוחר |
| רגישותם לערכי קייזן | רגישותם לערכי קייזן | רגישותם לערכי קייזן | רי' גישות רתק | רי' גישות רתק | רי' גישות רתק | רי' גישות רתק | יש אין | יש | יש רגישות גבואה | רגישותם לערכי קייזן |
| מהירות החישוב | מהירות | מהירות | מהירות | מהירות | מהירות | מהירות | אייטי אייטי | אייטי | אייטי | מהירות החישוב |
| סולם המידיה | שםי ומילה | שםי ומילה | שםי ומילה | שםי ומילה | שםי ומילה | שםי ומילה | ומעליה רוחים | ומעליה רוחים | ומעליה רוחים | סולם המידיה |
| משמעות להסקה | לא | לא | לא | לא | לא | לא | כן | | | משמעות להסקה |

3.4 שונות וסטיית תקן

אנחנו נשתמש בעיקר במשמעות ריבועיות, הידועה בשם: **שונות**.

עבור רשימת ערבים:

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

וְסִטְיַת הַתְּקֻנוֹ:

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

עבור טבלת 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 וסתירת התקין באוכולוסיה ובمدגים שונים זה מזה. באוכולוסיה, 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-1 – n ולא בn? נгла בהרצאה.

שימושים לממוצע וסטיית תקן: עבור עיקומת עמוקה, בערך 68% מהערכים הם במרחק של סטיית תקן אחת מהממוצע. בערך 95% מהערכים הם במרחק של שתי סטיות תקן מהממוצע.

חוק צביש: עבור כל התפלגויות,
לפחות 75% מהערכים הם במרחק 2 סטיות תקן מהממוצע.
לפחות 88.89% מהערכים הם במרחק 3 סטיות תקן מהממוצע,
באופן כלל לפחות $1 - \frac{1}{k^2}$ מהערכים הם במרחק k סטיות תקן מהממוצע.

3.5 ממוצע משוקל וסוגנות מצורפת

עבור k כיוות שוונוט, בהינתן N מס' התלמידים בשכבה מתקיים 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 כיוות שוונוט, השוונות 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 בביטויו הוא השוונות בין הקבוצות השונות, והחלק השמאלי היא השוונות בתוך הקבוצות (סוכמים).

תיקנו: למשל, סטודנט קיבל 70 בחשבונו ו57 בתנך. היכן הצליח יותר? בחישובו הממוצע היה 65 וסטיית תקן 3. בתנך 70 ו41 בהתאם. כיצד נדע? נורמל –

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

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

מכאן, נקבל שהתוצאות יהיו בתוך העקומת Z המפורסמת – התפלגות נורמלית סטנדרטית.

3.6 מdziי קשר בין מספר משתנים

למשל: האם יש קשר בין טמפרטורה ממוצעת באיזור לתנובת עצי פרי באיזור?
עד כה דנו במשתנה בודד, נדונו כתוב מס' משתנים.

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

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

מקדם המתאים של פירסום: ממד קשר הממלא אחר הדרישות הבאות:
א. ערכו המוחלט יהיה מקרים בלבד באשר הקשר מושלם (כל הנקודות על הישר - הקשר לינארי)
ב. סימנו של הממד שלילי או חיובי יבטא את כיוון הקשר (חיובי כאשר חיובי ולהפך).

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

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 מוגם אנחנו נחלק ב-1 – n .

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

4 הרצאה 4: הסקה סטטיסטית

4.1 מבוא

נשים לב, מה עשינו עד כה בקורס: שאלנו כיצד חוקרים מגלים? מנסחים השערה (הגדרת משתנים וסלומות מדידה), אוסף נתונים (אוכולוסיה ומודגס), מארגנים את הנתונים (מצגה טבלאית כמו שכיחות או ויזואלית כמו היסטוגרמה וכן מחשבים מודדים סטטיסטיים) ולבסוף מסיקים מסקנות. אם הנתונים נאספו על כלל האוכלוסייה אז סיימנו. **אם הנתונים נאספו על מוגם מייצג מוגם** **כל האוכלוסייה: עד לא סיימנו.**

נרצה לשאול כמה שאלות חשובות. האם ניתן להכליל ממדדיים במדגם למוגם באוכלוסייה? באיזו רמת בטחון ניתן לבצע חכללה זו? האם ניתן לקבל או לדוחות השערה ותחת אילו תנאים? המטריה העיקרית בהרצאה זו תהיה לבדוק – האם יש קשר בין תופעות המוגם לתופעות באוכלוסייה? כיצד נעשה זאת: באמצעות הסתברות. נראה כיצד הסתברות וסטטיסטיקה נפגשים.

4.2 הסקה סטטיסטית

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

הסקה דזוקטיבית: (כל \Leftarrow פרט), היסק לוגי, אמונות ההנחה מחייבת 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 הכל): הכללה, הנחות מובילות למסקנה בסבירות גבוהה. 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 הצדקה להסקה אינדוקטיבית במדוע (כל המדע מtabסס על הסקה שכזו)? לא נלמד זאת בקורס - זה מדעי הדשא. עם זאת, הבעיה הרכותית: כיצד לcame את מידת הוודאות שבתווך אי הוודאות? כן בקורס שלנו.

4.3 מושגים בסיסיים

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

תצפית: ווצאה של ניסוי מקרי מתוך המשתנה המקרי X .

דגם: ביצוע רצף תצפיות (ניסויים) X_1, \dots, X_N כאשר $X_i \sim F_{1 \leq i \leq N}$.

מבחן מקרי בגודל a מותך מ"מ X : מבחן של a 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 מתקאים $F \sim X_i$.

משמעות: דגימה מקרית (אקראית, רנדומית) עם החזרה של n איברים מותך אוכלוסייה עם תוכונה $X \sim F$ שקופה למבחן מקרי בגודל a מותך מ"מ מותאים $X \sim F$. ולהפוך.

מסקנה: מבחינה מעשית (\Leftarrow) נבעצ דגימה מקרית מותך אוכלוסייה גם כאשר בפועל נרצה לדוגום מ"מ. וכן מבחינה תאורטית (\Rightarrow) נוכל להשתמש בכל מה שהוא יודע על מ"מ על דגימה מקרית.

נשים לב: תכמה של האוכלוסייה נקראת פרמטר, וערכו קבוע אך לא בהכרח ידוע. מدد המבוסס על המדגם נקרא סטטיסטי וערכו ידוע אך לא בהכרח קבוע.
עבור אוכלוסייה בגודל k ניתן לייצר הרבה מדגמים שונים בגודל $k < n$ מכאן שככל סטטיסטי הוא משתנה מקרי עם התפלגות של עצמו - ההתפלגות הזאת נקראת התפלגות הדגימה של הסטטיסטי. (ככלומר, 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 באוכלוסייה, בסוג הסטטיסטי ובגודל המדגם. לכן נקבע כשנדבר על "התפלגות דגימה": התפלגות הדגימה של סטטיסטי מסויים s עבור מדגמים בגודל n שנלקחו מאוכלוסייה בה ערכי המשתנה מותפלגים לפי התפלגות F .

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

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

הוכחה:

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

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

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

הוכחה:

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

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

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

מזכירך להשתמש במדגם אחד? ברור כי במדגמים שונים עבור אותה אוכלוסייה יש ממוצעים שונים, אם נdagoms הרבה מדגמים ונחשב ממוצע לכל אחד, ממוצע הממצאים יתקרב מאוד לממוצע באוכלוסייה. אבל: אין בכוונתו לדגום הרבה מדגמים! אלא מדגם אחד וויאיד! אז: השאלה למשה: מהי הסבירות שהממצאים במדגם שדגמוני סיווה (בהרבה) מהממצאים באוכלוסייה? שאלה שוקלה: מהי הסבירות שהממצאים במדגם שדגמוני סוטה בהרבה מהתוחלת של הממם (ממוצע המדגם) עצמה? כלומר - כמה הערך של רוחק מהתוחלת של ממוצע המדגם. זו שאלה עדיפה לנו - כי אכן יש מדגם אחד כדי בדיק. לפיכך: נתענין במידת הפיזור של התפלגות הדגימה של ממוצע המדגם. סטיית התקן של ממוצע המדגם שווה לסטיית התקן של הממם המקורי מחלוקת בשורש n גודל מדגם ולכן: **ככל שהמדגים 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. لكن - נרצה מדגם יחיד גדול.

הוכחה:

לפי אי שוויון צביש'ב מותקיים

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

אם נפעילו על הממוצע \bar{X} קיבל

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

באשר n שואף לאינסוף, נראה כי ממוצע המדגם כלוא בין שני ערכי μ ולכן שווה לו.

נבחר $k \geq \varepsilon \frac{\sqrt{n}}{\sigma}$ ונצייבו, 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 תוצאות, כאשר מ"ס התוצאות שווה לאנוסף נקבל כי ההסתברות שסכום הממצאים שווה לתוחלת היא 1.

טענה: בדגימת מוגדים שגודלו n מתוקן ממ"ס המתפלג נורמלית עם תוחלת μ וסטיית תקן σ יהיה ממוצע המוגדים \bar{X} גם הוא ממ המתפלג נורמלית עם תוחלת μ וסטיית תקן $\frac{\sigma}{\sqrt{n}}$.

משפט הגבול המרבי: 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 . קלומר לכל $1 \leq i \leq n$ מתקיים $E[X_i] = \mu$. כך $S_n = X_1 + X_2 + \dots + X_n$ ש Ngd'ir

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

$$\text{מתקיים } E[Z_n] = 1 \text{ וכן } Var[Z_n] = 1. Z \sim N(0, 1)$$

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

הערה חשובה: מוגם "גדול מספיק" הוא זהה בגודל שנודלו ≥ 30 . ורק אם הוא בגודל שגדול מ-30 אפשר להשתמש במשפט הגבול המרבי.

5 הרזאה 5: אמידה סטטיסטית נקודתית

היכן אנחנו כתעת נמצאים? לומדים הסקה סטטיסטית. נושא זה מוחלך ל2: אמידת פרמטרים (הרזאה 5 – 6) ובדיקת השערות (הרזאה 9 – 7). אמידת פרמטרים מוחלכת ל2: בהרזאה זו נדבר על אמידה סטטיסטית נקודתית ובהרזאה הבאה נדבר על אמידה מרוחה בטוחן. היום נדבר על השאלה הבאה: כיצד והאם ניתן להכליל ממצאים במדגים לממצאים באוכלוסייה?

פרמטר: גודל קבוע המאפיין את כל האוכלוסייה.

סטטיסטי: ערך המוחושב ע"פ המדגם.

אמידה היא הערכת (שערון) ערך הפרמטר ע"פ סטטיסטי המדגם.

5.1 אמידה סטטיסטית

ישנן שתי שיטות לעריכת אמידה סטטיסטית.

1. **אמידה נקודתית:** הכניסה המומוצעת של משפחה בת 4 נפשות היא 11,500 שקלים בחודש -

על סמך המדגם מחושב סטטיסטי אחד.

2. **רווח סטטיסטי:** בהסתברות של 80% ההוצאה המומוצעת של משפחה בת 4 נפשות בישראל היא בין 8000 ש"ח ל-16,000 ש"ח - על סמך המדגם מחושב טוחן של ערכיהם.

لامידה סטטיסטית נקודתית ישנן בעיות:

א. בעיה מוחותית - הסטטיסטי הוא רק אומדן. כיצד נדע את ערך הפרמטר ביחס לאוכלוסייה כולה? (לא נדע). מדווקא מותר להשתמש בהנחה אינדוקטיבית? (לא בקורס זהה).

ב. בעיה מעשית - בעיה כמותית, באיזה סטטיסטי כדי לי להשתמש כדי לאמדוד משתנה מסוים? באיזה אמדים קיימים ואיזה תוכנות יש להם? מה נדרש לאמד טוב?

5.2 בעיית האמידה

נתון: עברו משתנה מקרי $X \sim F$ ונתנו מדגם מקרי x_1, x_2, \dots, x_n ב"ת באשר $x_i \sim F$ ($\forall 1 \leq i \leq n$)

הנחת עבודה: אנו ידעים את צורת התפלגות של X - פונקציית ההסתברות או הצפיפות אך לא ידעים את הפרמטר.

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

דוגמה: אמידת זמן חיים של נורה (λ) $X \sim \exp(\lambda)$. אמידת פרמטרים של התפלגות נורמלית גבוהה או משקל של בניים או בנות $X \sim N(\mu, \sigma^2)$.

טרמינולוגיה:

מעבר פרמטר באוכלוסייה θ נסמן את האמד בمدגם $\hat{\theta}$.

דוגמה: עברו התוחלת μ נבחר אמד שיחיה המומוצע: $\sum_{i=1}^n x_i = \frac{1}{n} \hat{\mu}$. (מטרת השיעור היא להסביר מדוע בהכרח המומוצע היא אכן טוב לתוחלת. זה מוכיח שהוא והאמד אכן 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 ולכך מחשבים את האמד $\hat{\mu}$.

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

הגדרה: נתנו מדגם מקרי X_1, \dots, X_n . אנו רוצים לאמוד את ערכי θ מתוך המדגם. אז,

1. פרמטר - פונקציה של ערכי האוכלוסייה. יכול להיות תלואה בפרמטרים לא ידועים.

2. סטטיסטי-פונקציה של ערכי המדגם. אינה תלואה בפרמטרים לא ידועים.

דוגמה: $\sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^n x_i$ (השוניות) אינו סטטיסטי כי תלואה בענ.

3. אמד (*estimator*) - סטטיסטי שבעזרתו אומדים פרמטר בלתי ידוע (פונקציה כללית). לדוגמה: הממוצע הוא אמד לתוחלת. **שאנו נוחשים אמד:** אנחנו מוחשים נסחה.

4. אומדן - המספר עצמו שמציבים בנסחה (אמד) עברו מקרה ספציפי. התוצאה שקיבלו עבור האמד במדגם ספציפי (תוצאה ספציפית).

דוגמה: הממוצע במדגם הטלת קופיה 1, 2, 1, 4, 2, 6, 4, 2, 5 הוא האומדן במדגם:

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

5. שגיאת האמידה - המרחק בין ערך האמד לערך הפרמטר: $\theta - \hat{\theta}$. נשים לב כי את θ איננו יודעים. אז כיצד יעזר לנו ליחס ערך אמד (במקרה 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 גדר רשמי של השגיאה של אמד הינה:

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

5.3 תכונות של אמידים

מהו אמד שהוא טוב?

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

תכונות ממוצע המוגם:

- עבור תוחלת μ נגדיר את ממוצע המוגם כאמד: $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$ האם הממוצע של המוגם הוא אמד טוב לתוחלת? א. אכן אמד עקי - ככל שהמוגם גדול, ערך האמד \bar{X} מתכנס לערך הפרמטר באוכטוסייה. זה מוגע בבדיקה מחוק המספרים השלמים:

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

- ב. אכן חסר הטיה - ראיינו כי $E[\bar{X}] = E[X_i]$ בהערכתה הקודמת (באשר $E[X_i] = \text{Bias}(\hat{\theta}, \theta)$ מוגם כלשהו), ומכאן שנקבע כי אכן $E[\bar{X}] = \mu$

- טענה (עבור כל התפלגות):** בהינתן מוגם מקורי x_1, x_2, \dots, x_n ב"ת מתוך מ"מ X עם תוחלת μ ושונות σ^2
- א. אמד לתוחלת שהוא חסר הטיה הוא הממוצע $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^n x_i}{n}$
- ב. אמד לשונות (בהינתן שתוחלת 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':

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

כעת נדונ באמוד 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 $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$

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 מדים (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.
 $V(\theta_2) < V(\theta_1)$ את θ_2 המקיים

יעילות של אמדים: במקרה הכללי - תוחלת ריבועי השגיאות הינה

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

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

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

5.4 שיטות אמידה

5.4.1 שיטת המומנטים

שיטת המומנטים היא שיטת אמידה על פי פרמטרים המאפיינים התפלגות של אוכלוסייה מסוימת. נניח משתנה מקרי המתפלג F עבורה ישם k פרמטרים בלתי ידועים נגדיר **פונקציה מייצרת מומנטים** (mfg). נאמוד את המומנט μ_k באמצעות ממוצע חזקה 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 האنبות" וככז זה ממשיך.

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

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

שיטת אומרת 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.

דוגמא:

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

הספק לנו מומנט ראשון כי רצינו למצוא משתנה יחיד. נתבונן בדוגמה נוספת:

נניח משתנה מתפלג 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 קיבלנו שתי משוואות בשני נעלמים. הרו x נתונים לנו וגם a . קיבל

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

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

וכך, מנתונים שידוע שמתפלגים בצורה איחוד, הצלחנו למצוא אמד a ו b הנדרשים.

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

5.4.2 שיטת הנראות המרבית

נניח שהטלתי מטבע מס' פעמיים. בכל ההצלחות קיברתי 5 (זה ניסוי ברנולי). לפי מה שזה נראה - נראה כאילו $1 - p$. נגידיש: **נראה**.

פונקציית הנראות L : בהינתן ערך p ניתן לחשב את פונקציית הנראות. נראה כי בהינתן משתנים ביןומית, נניח ואנחנו יודעים כי ההסתברות שיצא 7 פעמים 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 לפ p ונשווה לאפס.

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

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

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

נשים לב. מדוע המרנו LnL ? הרבה יותר קל locator כז.
 שנייה: קיבלנו הוכחה מעניינת לכך **שהשכיחות היחסית היא אמד נראות מרבית עבור פרמטר p בתפלגות ביןומית**.

להלן השיטה:
 א. נגדיר 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. נגדיר את לוג פונקציית הנראות

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

- ג. נגזר את LL לפי θ ונשווה לאפס למציאת ערך קיצון.
ד. נגזר את LL פעמייה לודא מקסימום.

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

מדוע אן"ם טוב לנו?

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

| לשונן: | | | | |
|---------------------------|---|-----------------------------|------------------------------------|--|
| האם חסר-הטיה – אוח"ה | האם גראות מירבית – אנ"ם | התפלגות מ"מ | מודל תאורטי | |
| $ x $ | $= \bar{X}/n$ | $X \sim \text{Bin}(p)$ | בינוי | |
| $\max\{X_1, \dots, X_n\}$ | | $X \sim U(1, b)$ | אחדידה | |
| $ x $ | $= \bar{X}$ | $X \sim P(\lambda)$ | פואוטוני | |
| $\max\{X_1, \dots, X_n\}$ | $= 1/\bar{X}$ | $X \sim G(p)$ | אגומטרי | |
| $ x $ | $= 1/\bar{X}$ $= 1/\bar{X}$ | $X \sim \text{Exp}(\theta)$ | מעריצי | |
| $\max\{X_1, \dots, X_n\}$ | $= \bar{X}$ $= \sum_i (X_i - \bar{X})^2 / n$ | $X \sim N(\mu, \sigma^2)$ | נורמלית: - תחלפת - - שונות - | |

6 הרצתה 6: אמידה סטטיסטית של מרוחקי בטחון

היכן אנחנו? לומדים תהליכי ייסוי, אנו בחלק של אמידת פרמטרים + בדיקת השערות, נושא לו קראנו סטטיסטיקה היסקטית.

- ראינו כי סטטיסטיקה היסקטית מתחולקת ל-2:
א. אמידת פרמטרים: אמידה נקודתית (שיטות המומנטים ושיטות הנראות המרבית) ומרוחקי בטחון
- נושא ההרצאה הנוכחיית.
ב. בדיקת השערות - בהמשך.

6.1 רוח סמך של ממוצע המדגם

נتبונן בממוצע המדגם \bar{X} כאמד נקודתי ל佗לה μ . שגיאת האميدה של ממוצע המדגם היא $\mu - \bar{X}$. נשים לב כי שניית האמידה לא ידועה לנו, ושגיאת האמידה היא משתנה מקרי בעצמה. תחת תנאים אלו נרצה לבדוק את דיוק האמד. דיוק האמד לא יכול להיות מודר אבסוטלי - אלא מודר הסתברותי. לכן נshall: מהי ההסתברות לכך ששגיאת האמידה של האמד תהיה גדולה מ(טוחן)? נאכר כי לכל $n \geq 30$ עבור משתנה מקרי X בעל תוחלת μ ושונות σ^2 מתקיים $\bar{X} \sim N(\mu, \sigma^2)$ (משפט הגבול המרכזי).

דוגמה.

בمدגם שגודלו 25 מתוך מ"מ X המתפלג נורמלית בעל סטיית תקן $\sigma = 10$ ו佗לה μ לא ידועה, מהי ההסתברות שממוצע המדגם יהיה שונה מה佗לה בלא יותר מ-4 יחידות? נראה כי נתון $X \sim N(\mu, 100)$ וכן $n = 25$, ולכן $\bar{X} \sim N(\mu, 4)$ (שונה ב-4 יחידות = השונות היא 4). קלומר, נדרוש

$$P(|\bar{X} - \mu| < 4) = P(\mu - 4 < \bar{X} < \mu + 4) =_{Z = \frac{\bar{X} - \mu}{\sigma}} P\left(\frac{\mu - 4 - \mu}{2} < Z < \frac{\mu + 4 - \mu}{2}\right) =$$

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

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

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

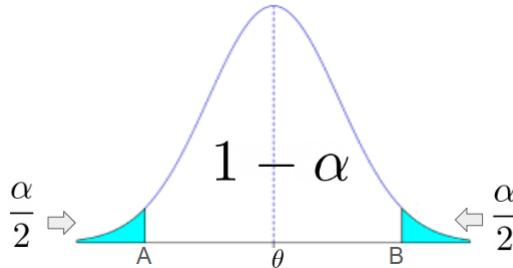
ישנה שיקילות ב כלל הערך המוחלט. המשמעות לשינוי זה היא - במדגם שגודלו $n = 25$ מתוך מ"מ המתפלג נורמלית עם סטיית תקן 4 ו佗לה μ לא ידועה, בהסתברות 0.95 הרוחה יכול את μ . קלומר - מצאנו אינפורמציה חשובה באשר ל μ . נראה לביטוי זה: **רוח סמך של μ ברמה של 0.95**.

6.2 רוח סמך - הגדלה פורמלית

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

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

כך זה נראה בגרף:



הטעות האפשרית - היא α (הטוריקי) ורמת הסמך היא החלק הלבן.

6.3 רוח סמך לממציע עבור התפלגות נורמלית כאשר השונות ידועה באמד לתוחלת

עבור X בעל תוחלת μ , שונות σ^2 וגודל מוגם $n \geq 30$, 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 בשביל שאנו ימין (הסתברות) $P(-z < X_Z < z)$ רוח סמך בrama 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 $z_{0.025} = 1 - 0.025 = 0.9750$ ונקבל $\frac{\alpha}{2} = 0.025 = 0.025$, נלכ' לחפש ערך זה (0.9750) בטבלת התפלגות הנורמלית Z . נראה כי 1.96 מניב הסתברות זו כלומר $Z = 1.96$ ולכן רוח סמך של 95% הוא כאשר $Z = 1.96$. אם נרצה רוח סמך של 90% כלומר $\alpha = 0.1$ ולבן $z_{\frac{\alpha}{2}} = z_{0.05} = 1 - 0.05 = 0.95$ וערך זה מתקבל עבור $Z = 1.645$

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

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

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

דוגמה 2. אורך החיים של נורות מותוצרת מפעל הניצץ מתפלג נורמלית עם תוחלת μ וסטיית תקן $\sigma = 22$.
נדגמו רנדומית 16 נורות ונמצא שאורך החיים הממוצע הוא 863. מצא רוחם בסמך 90%.

פתרונות: נראה כי מתקיים $\alpha = 0.1$ ולכן $z_{1-\frac{0.1}{2}} = z_{0.95} = 1.645$ ואם נציב בנוסחת רוחם 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 רוחם סמך לממוצע כאמד לתוחלת באשר השונות אינה ידועה

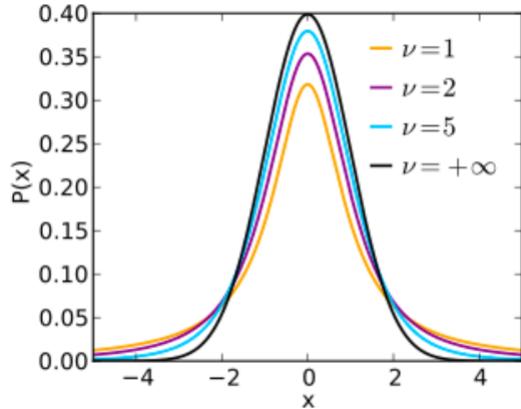
נשים לב כי ברוב המקרים השונות לא ידועה לנו מראש. לכן נצטרך לאמוד את השונות σ^2 בעזרת אמד חסר הטיה, כפי שראינו בהרצאה 5.

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

כעת, שינינו את התפלגות הדגימה. כיצד יתפלג המ"מ החדש? $Z_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{\hat{S}}{\sqrt{n}}} \sim N(0, 1)$ עבור מוגדים $n \geq 30$.
לכן, רוחם הסמך ברמת סמך (רמת בטיחון) של $\alpha - 1$ אחוז עבור μ כאשר השונות אינה ידועה עבור מוגדים $n \geq 30$ 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 למוגדים קטנים? עבור מוגדים קטנים $n < 30$ נסמן $t(v)$ ה**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 שהרץ יהיה אך האחרון חייב להתאים את הערך הכלול לממוצע. 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, כיוון שתלויים עוד ווד. אנו נסמכים על חישוב הממוצע, וכן דרגת החופש v היא גודל המדגים פחות אחד. ככלומר,

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

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

דוגמא. במדגם בגודל $n = 9$ מאוכטוסייה מתפלגת נורמלית נמצאה הממוצע 114. האומדן לסטטיסטיקת התקן באוכטוסייה הינו 12. מצא רוח סמך לתוחלת ברמת סמך של 95%.
פתרונות: נתון לנו $n = 9 < 30$ ולכן נשתמש בטבלת התפלגות t . וכן $v = 9 - 1 = 8$ ו- $\bar{X} = 114$. מכאן נקבל כמו כן $\alpha = 0.05$ וכן $t_{\frac{\alpha}{2}}(8) = t_{0.025}(8) = 2.306$.

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

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**:

$$\hat{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}} < \mu < \hat{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{S}}{\sqrt{n}}$$
- 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 הטיעות הינו $ME = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ ומתקיים $ME < \mu < \bar{X} + ME$.

7 הרצאה 8 + 7: הסקת סטטיסטית - בדיקת השערות

בהרצאה הקרובה נדון על בדיקת השערות במשתנה אחד וקבוצות קטגוריאליות: הרעיון הכללי וריאציות שונות על אותו נושא - משתנים מסוימים שונים, השוואות שונות.

7.1 השערת האפס וההשערה האלטרנטיבית

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

נסמן את ההשערה הראשונית - **השערת האפס ב** H_0 , ואת **ההשערה האלטרנטיבית 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 גובהם הממוצע של גברים אסיאתיים גדול משל גברים באוכלוסייה?" אזי אם נסמן μ_{Asian} כממוצע הגובה של גברים אסיאתיים ו $\mu_{Non-Asian}$ כממוצע הגובה של גברים שאינם אסיאתיים אזי:

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

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

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

איך נחליט איזו מההיפותזות נכונה?

קיימות שתי גישות בסיסיות לקבלת החלטה:

- חישוב הסבירות של השערת האפס H_0
- גישה קבלת החלטות (מעורר השגיאה).

נשים לב: ההחלטה יכולה להיות מוטעית, אך נשתדיל כי הסיכוי לכך יהיה קטן ככל הניתן. וכן: אנו רק מנסים לשלול את השערת האפס - לא להוכיח שההשערה האלטרנטיבית 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 **אין** 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 | שגיאה מסוג II החלטה נכונה (True Negative) | החלפת הבדיקה (על סמך המדגם) H_0 שגיה במציאות |
| דחיתת H_0 | אי-דחיתת H_0 | החלטה נכונה | החלטה נכוןת H_0 |

7.3 חישוב הסבירות של השערת האפס

נניח שאנו רוצים לבדוק אם ממוצע המדגם של גברים (\bar{X}) שונה מהממוצע של האוכלוסייה (μ) בגברים. 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:

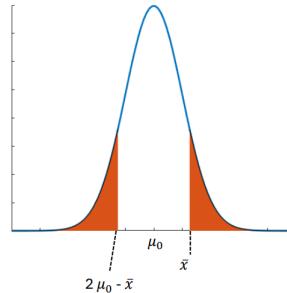
$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

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.odal זה מכונה p והוא הסוג הראשון של השגיאה.
- מה ניתן לומר עליו?
- א. $0 \leq P \leq 1$.
- ב. חסר'ויות
- 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 מזו שנקצתה היא $p - Value$). $p - Value$ (כלומר, ככל שההפס נכונה) הסתברות ש- \bar{x} גדול או שווה מזו שנקצתה היא $p - Value$. ככל יותר קטן, פחות סביר ש- H_0 נכונה.



צדדים ב- $P - value$

- א. מבחן *right-tailed*: אם אנחנו חושדים מראש (לפניהם שראינו את הנתונים) כי $\mu < \mu_0$.
- ב. מבחן *left-tailed*: אם אנחנו חושדים מראש (לפניהם שראינו את הנתונים) כי $\mu > \mu_0$.
- 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 - | $C = \mu + Z_a \cdot \frac{\sigma}{\sqrt{n}}$ | $C^+ = \mu + Z_{a/2} \cdot \frac{\sigma}{\sqrt{n}}$ $C^- = \mu - Z_{a/2} \cdot \frac{\sigma}{\sqrt{n}}$ | $C = \mu - Z_a \cdot \frac{\sigma}{\sqrt{n}}$ |
| 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 לקבל או לדחות את השערת האפס!!!

7.4 שלבים להחלטה אם לקבל או לדחות את השערת האפס

- כדי להחליט אם לקבל או לדחות את השערת האפס יש לבצע את הצעדים הבאים:
- א. להחליט לפני החישוב, על סף שאמם $P - value$ – 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. לדחות 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 מראש. האם זה גורר ש- H_0 נכון? לא! זה רק אומר שאי אפשר לדחות 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 לדחות 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* יש לקבוע את הסף $P - value$ ל- $\frac{\alpha}{n}$ באשר α הוא הסף למבחן בודד, ו n הוא מס' המבחןים. תיקון זה הוא שמרני, ישןם 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 המבחן הדורש

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 משמעותי 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 בחשבון 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 שנינו *FN* – 1, כלומר: המשלים של הסיכוי לשגיאה מסווג 2 – ההסתברות לקבל את השערת האפס בטעות (β). לעומת זאת, העוצמת המבחן היא ההסתברות לדוחות את השערת האפס כשלעצמה.

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

كيف מחשבים את עוצמת המבחן? ראשית מחשבים את $\phi(\frac{C-\mu_1}{\sigma/\sqrt{n}})$ – בעוד $C = \beta$ והערך הקרייטי (כתלי בבדיקה שקבע מתי נדחה את H_0 לפי רמת מובהקות α), μ_1 הוא הממוצע האמיטי שלכארהה נכון 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 שאותו מנסים לאות (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 שמדובר מוחפשים אם הוא אמיתי 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 הטעאה לא מושלם, קו החחלטה יכול לטעת. נתבונן בשני עולמות מקבילים:

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

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

7.10 מבחני השערות במדגמים 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 שכזה: 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 התובאה 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 מקרהבודד 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 "כמה ייחדות רעש אני רחוק מהממוצע המוצהר? 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 הזרב אינו מתקיים.

דוגמה לממוצעים. טענה: H_0 : הגובה הממוצע של גברים בישראל $= 175 \mu$. 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 מקובל $= 173 \mu$. 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 175μ ? 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 ממשמעותי? כיון שמדובר במשתנה בינומי, $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$.

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

ובcut, אם נרצה רמת מובהקות של α : אזי לא 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 הקרוبات, $p = 0.4$ 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 $Z = 2.74 > 1.96$. cut, $2.74 > 1.96$ ברמת מובהקות של 5% ולכן דוחים את H_0 ברמת מובהקות 5%.

חשיבות: אם מתקיים $10 > np > n(1-p)$ אזי אפשר לומר כי פורפורציית המדגם מתפלגת נורמלית ככלומר ($\hat{p} \sim N(p, \frac{p(1-p)}{n})$ דהיינו $\hat{p} = p + \frac{p(1-p)}{n}$). כעת יוכל להשתמש ב- Z בשיביל לדעת את פונקציית הצפיפות המצטברת של משתנה 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 קבוצות שונות? נניח כי הקבוצות בלתי תלויות זו בזו. כמובן: בדקנו שתי קבוצות, אסיאטים ולא אסיאטים. הממוצע של האסיאטים בגובהם הינו 174 ס"מ והממוצע של הלא אסיאטים בגובהם הינו 177 ס"מ. אנו שואלים - האם ההפרש של 3 בנים היה מקרי, או שפושט 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 $x_1 = 150$ לחסוך עליה. איז $P_1 = \frac{150}{500} = 0.3$ ליחסו עליה. וכן $n_2 = 600$ נשים ראו פרסום $x_2 = 120$ ליחסו עליה איז $P_2 = \frac{120}{600} = 0.2$. כיצד נבדוק את ההשערות? השערת האפס הינה $P_1 = P_2$ והאלטרנטיבית היא שהם שונים. לפיכך, נחשב את Z לפי $p = \frac{x_1+x_2}{n_1+n_2} = 0.245$. מכאן נחשב את סטיית התקן שתצא $\sigma_{P1-P2} = 0.026$. נחשב את Z לפי הנוסחה מלמעלה 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 מותוך המדגם S . אם $30 < n$ נוכל להשתמש ב t , ואם $n \leq 30$ ניהיה t .
 חייבים להשתמש בטבלה.

لسיכום: מתי נשימוש ב מבחן זוגות? נשימוש 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 מבחנים אפרמטריים: מבחן χ^2

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

עד כהבדקנו עם מבחנים פרמטריים, הנקנו הנחות על הנתונים - הנתונים מתפלגים נורמלית, התוחלת היא μ , סטטיסטית התקן היא S וכדומה. מבחני Z ו- t הם מבחנים פרמטריים. אם כן, אלו מבחנים שרגישים הרבה יותר לנתחונים חריגים, דורשים שהנתונים יהיו נורמליים ולא עובדים טוב על נתונים קטגוריאליים.

מבחנים אפרמטריים הם מבחנים שאין מניחים הנחות על התפלגות הנתונים. לכן: פחות רגושים לנתחונים יוצאי דופן, אין צורך לבדוק את התפלגות הנתונים והם עובדים הן על נתונים אודינליים והן על נתונים קטגוריאליים. מה החסרונות? הם חזקים פחות מ מבחנים פרמטריים - דורשים מדגם גדול יותר.

מבחן χ^2 הוא מבחן אפרמטרי שוענה על השאלה הבאה: האם ההתפלגות שאני רואה בנתונים תואמת להתפלגות שציפיתי לה? למשל: זריקת קובייה - יש לך קובייה ואתה זורק אותה 600 פעמים. אם הקובייה הוגנת, אתה מצפה לראות בממוצע כל מספר 100 פעמים. אם בפועל קיבלת רק 107 עבור 1 ו-93 עבור 2 למשל, האם הבדל זה קרה בנסיבות במקורה רעש או בכוונה והקובייה אינה הוגנת?

הרעין זה, מודיעד כמה רוחקה ההתפלגות שראינו מהתפלגות שציפינו. לשם כך, נتابعו בנוסחה הבאה:

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

באשר, O_i - מה שראינו, המשתנה שנצפה *Observed*. וכן E_i זה המשתנה שציפינו אליו - *Expected*. במקרים אחרים הנוסחה אומרת: חשב את הפער, חלקי מה שציפינו אליו. מודיעד מעלים בריבוע? בשביל שכיוון הפער לא יהיה חשוב. מודיעד מחלוקת ב- E_i ? בשביל ליצור הפרש יחס. אם ההפרש יצא 20 אך קיבלנו 20 וציפינו ל-40, זה לא אותו דבר כמו שציפינו ל-120 וקיבלו 100.

לוקחים דוגמה של קובייה למשל, עם נתונים ידועים. מחשבים ומתקבלים כי $\chi^2 = 2.58$. מס' דרגות החופש הינו $1 - k$, כיון שישנם 6 ערכים אפשריים בקובייה) אזי מס' דרגות החופש הינו 5. שכן, אם קבענו 5 ערכים עבור כמה הטלות יצא בהם 1, 2, 3, 4, 5. איזי ברור שכיוון שסכום הטילות הוא 600 ידועה התוצאות לאחרונה. כתע, מחפשים בטבלת χ^2 את $\alpha = 0.05$ (כבות ודאות) וכן מס' דרגות חופש $5 = u$. מקבלים כי $C = 11.07$ (ערך קרייטי). כיון שאצלנו וכן $\chi^2 = 2.58 < 11.07$ אנחנו לא דוחים את H_0 והקובייה הינה הוגנת.

באופן כללי עבור מבחן χ^2 : אם קיבל $\chi^2 > C$ אז לא נדחה את H_0 ואם קיבל $\chi^2 < C$ כן נדחה את H_0 בمبرח ההשערות. וכן, אם χ^2 קטן איזי זה קרוב למה שציפינו, יכול להיות רעש רגיל ולפנן לא דוחים. אם χ^2 גדול איזי זה רחוק ממה שציפינו - לא ניתן שזה רק רעש, ולפנן דוחים.

הערה חשובה. נבחן כי χ^2 בהכרח חיובי, לכן אם $0 = \chi^2$ זו התאמה מושלמת, בדיק מה שרצינו. בכוונה מדבר 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 Mann

מודיע אנחנו זוקרים לבחן נוסף? עד כה, בשביל להשוות שתי קבוצות השווינו ממוצעים עם מבחן t .

אבל מבחן t דוחה הhipothesis נורמלית, מקרים גדולים ונתונים רציפים. 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.

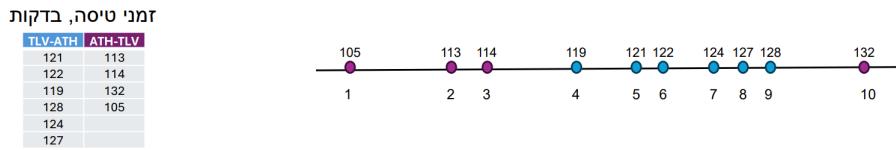
המודגש בודק האם החזיניות של שני התפלגות זהים. משווים את החזיניות של שתי ההתפלגות. נתנות דגימות x, y מהתפלגות אוריינטליות רציפות F_x, F_y בהתאם. כיוון שההתפלגות רציפות $P(x = y) = 0$:

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

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

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 ממיטסota מאטונה לישראל?



המבחן מציע את הרעיון הבא: נסדר את הנתונים על קו ישר" באשר נזכיר לכל נקודה מהין היא הגעה - מאיו דגימה וכן נסדרם בסדר עולה. נבחן כי אם H_0 נכון (אין הבדל) אז הם צריכים להיות מעורבים (כי החזינים 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 ערךים 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 ($R'_1 < R_1 = 16$). וכך זה לא מובהק סטטיסטי - זה לא דוחה את השערת האפס. אם היינו מקבלים ממש הפרומות מוגבלות 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 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 קطن יכולנו לחשב את מספר האפשרויות לסדר, אך במקרים עם n גדול זה מספר אסטרטוני. לפיכך, כשמדובר במספיק גדולים $20 - 10 > n_1, n_2 > 10$ נשתמש ב מבחן Z וגיל באשר החישוב יהיה כך:

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

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 זהים הם יקבלו את אותו ערך דירוג. הערכה שנייה. علينا לחשב ערכיו 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? לקבוצות בלתי תלויות ישנו מבחן פרמטרי של Mann – Whitney – U . לקבוצות 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 (תאומיים, אותו אדם), הנתונים לא נורמליים או יש חריגות, מדים קבועים. נרצה לבדוק אם החץון של x_i שונה מהחץון של y_i .

לדוגמה, האם מרתון בוסטון בשנת 2024 היה מהיר יותר ממרתון 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 מחזירים את הסימן שלו ומתקבלים סידור שלהם באופן מדורג. מחשבים את w^+ ו- w^- : שתי קבוצות המייצגות בהתאם את האינדקסים המדורגים של האיברים סימנים חיובי, ואי-בריים שסימנים שלילי. מתחבוננים את 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:

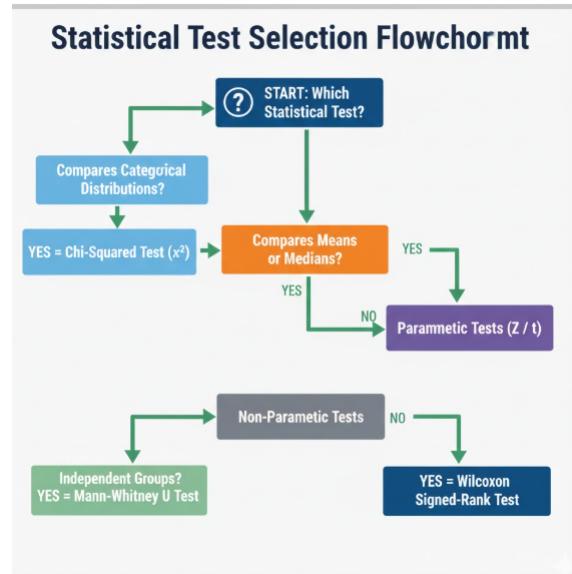
$$\mu_W = \frac{n(n+1)}{4}, \sigma_W = \sqrt{\frac{n(n+1)(2n+1)}{24}}, Z = \frac{W - \mu_W}{\sigma_W}$$

לסיכום, שלבי המבחן:

- א. חשב את ההפרשים (X_i, Y_i) $D_i = X_i - Y_i$ לכל זוג (X_i, Y_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?



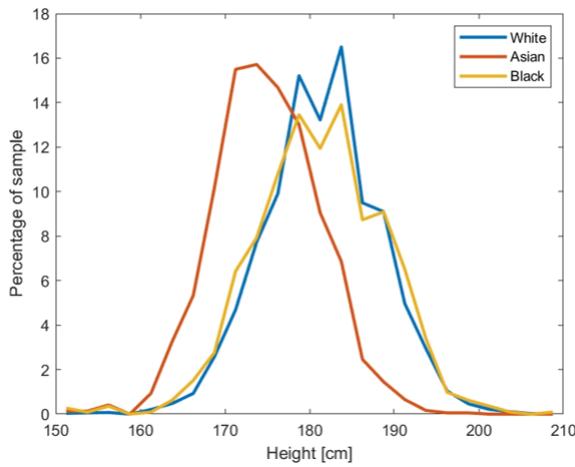
הרצאה 9 : 8 הרצאה (Analysis of Variance) Anova

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 עושים זאת במשתנה אחד ובהמשך נרחיב ליותר משתנות אחת 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 ידוע הרבה קבוצות נוספת: אנשי לבנים, אסיאטים, שחורים, יהודים, אינדיאנים וכו' הלאה. כתע - ניתן לחלק את התנאים ליותר מ-2 קבוצות ולשאול שאלה אחרת: האם גברים שונים בגובהם כתלות בקבוצה האתנית שאליה הם משתייכים?

תבונן בהטפלות 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.

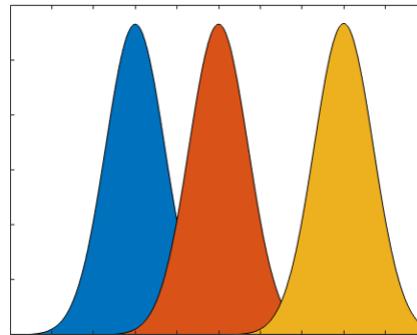
8.1 הנחות יסוד לבחן אנוּבָה (בסיסי)

- 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 זהה.
 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 שהגושים אינטגרליים יהיו יותר צפופים אחד מוהני ופחות מעורבבים אחד בין השני - ואם זה יהיה נכון: נוכל 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:

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

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

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 אסייתים + 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: $\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{i,j}$ באשר $1 \leq j \leq q$.
 ב. הממוצע של כלל 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 הוסיףו לעמוי הרגנה):

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

וינו הצד השני:

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

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

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

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

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

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

ושוב, מתקיים כי s_j הוא משערך השונות של הקבוצה j .
קיבלנו 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 משתי 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:
הפיורר סביב הממוצע של כל קבוצה:

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

הפיורר (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}

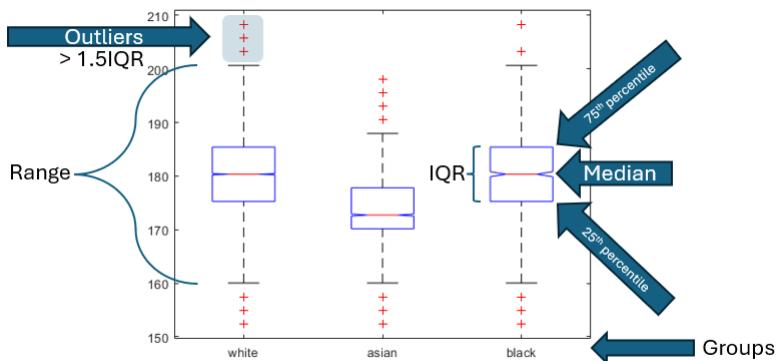
כפי שאמרנו, אנו רוצים לבדוק את היחס בין פייזור הנקודות בתוך הקבוצות לבין פייזור הנקודות בין הקבוצות. לכן, המדר הסטטיסטי שנבחן הוא:

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

כל ש F גדול, אז קל יותר לדוחות את השערת האפס. מדוע? אם F גדול, אז המונה גדול, כלומר S_{Between}^2 גדול והמומיצים 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 שלROLONTITAT לנו נמצא שטח 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 - עברנו ל מבחנים 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 ב מבחןanova, 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 בוצע בדיקת השערת בין כל זוג. כמובן, ניבור על כל הזוגות האפשריים. נזכר - כי אנחנוCutת עושים 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 למניע 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 שלהם על הקלט והנתונים.

- נדון ב מבחן 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 m דוגמאות k 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, $HSD = \frac{|\mu_A - \mu_B|}{\sqrt{\frac{s^2}{n}}}$ הוא ההבדל המינימלי בין זוג ממוצעים שייחס מובהק סטטיסטית (באוטו ס' בו חושב q) - ככלומר HSD יהיה הס' שבו אנחנו נגיד - 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- $\approx \sqrt{2}z \approx \sqrt{2}Z$ וכאן לרוב נלך 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 $ANOVA$ - 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 $k - 1$ 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 הס' 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 | בדד סטטיסטי | סכום הריבועים | דרגות חופש |
|--------------|-------------------------------|-----------------|------------|
| בין קבוצות 1 | $SS_{Between1}/SS_{Residual}$ | $SS_{Between1}$ | a-1 |
| בין קבוצות 2 | $SS_{Between2}/SS_{Residual}$ | $SS_{Between2}$ | b-1 |
| בתוך הקבוצות | | $SS_{Residual}$ | (a-1)(b-1) |
| 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 |

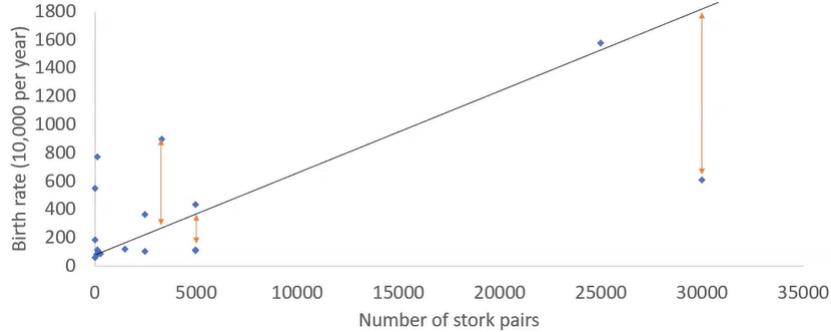
מה עושים אם הנתונים לא מפולגים נורמלית? ככלומר - אנחנו ביוטר ממשטנה אחד, האם יש מבוזן דמי קורסקל? יש, הוא לא מוכיח מוצלח. לכן: לא נלמד אותה. ולכן מסקנה: אם הנתונים לא מפולגים נורמלית - אין לנו מבחן טוב עבור יותר ממשטנה אחד.

9 הרצתה 10: רגסיה לינארית

מדובר במשפחחה אחרונה של בדיקת אחרונות שנלמד, כתע אנחנו למדים ללמידה 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...). כתע 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 במדינת מסויימת (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 מדים 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 במדד 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? נזoor לפי b , w ונשווה לאפס:

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

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

$$\begin{aligned} \cdot \sum y_i x_i &= w \sum_i x_i^2 + b \sum x_i \\ \text{מהמשוואה השנייה קיבל כי} \quad \sum y_i &= w \sum x_i + nb \\ b = \frac{1}{n} \sum y_i - \frac{1}{n} w \sum x_i &\implies b = \bar{y} - \bar{w}\bar{x} \end{aligned}$$

נזהoor למשוואת הראשונה,

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

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

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

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

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 נגור 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 שמספרים 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 $W^T X^T XW$ יגרז להיות $2X^T XW$ לפי הכלל הזה - $2Az = \vartheta(z^T Az)$). 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 הגודל שלה: די קטן. נעיר כי אי אפשר לפתחו לפי חוקי מטריצות, שכן X אינה מטריצה ריבועית אך $X^T X$ כן כן.

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

מכאן, המשמעות הנ"ל נותנת לנו נוסחה סגורה לחישוב קו הרגרסיה W . מדובר כמובן בתחום שימוש באלגברה ליניארית בשbill לפטור את מודל הרגרסיה (ולא במשמעות הראשונית?) שכן: בכל מצב שיש יותר ממשתנה בלתי תלוי אחד, המשוואות רק מתחילה להסתבך. ולכן: תמיד (!) להשתמש בשיטה זו.

לבסוף, לאחר שה חישבנו את W קיבל את קו הרגרסיה. זה, יש קשרلينו את קו הרגרסיה $y = 0.0287 + 233.76 x$ - זו התאמת טובת? 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 $-1 < r = 1$, 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 :

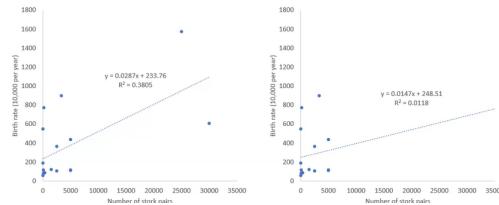
$$r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} = \frac{w^2 \sigma_x^2}{\sigma_y^2} = \frac{w^2 E(x - \bar{x})^2}{\sigma_y^2} = \frac{E((b + wx - \bar{y})^2)}{\sigma_y^2} = \frac{E(\hat{y} - \bar{y})^2}{E(y - \bar{y})^2} = \frac{SS_{REG}}{SS_{TOT}}$$

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 הגיעה מאוד לנוטים חריגים. אם יש לנו שני נתונים חריגים מאוד מהשאר - הם מאוד מושפעים 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 משתנה אחד? במקרה זה אי אפשר לציר גורם עם שיפוע Ci יsono יותר משתנה 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 מפותח 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 x_i שווה לאפס. והאלטרנטיבית: הוא שונה מאפס.

כיצד נעשה זאת? נסתכל על הסטטיסטיק $\frac{\hat{w}_i - 0}{SE(w_i)}$.
הויראנס של השארית ϵ מוחושב ע"י $\frac{\|y - XW\|^2}{n-k} = \frac{RSS}{n-k}$ (כיוון שהראש ב"ת בכל המדידות ניתן כתוב כי $\epsilon \sim N(0, \sigma^2 I)$ לשורוך 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). ככלומר, מסתכלים על ערך זה בהסתפלות t ועושים מבחן t עם 1 דרגות חופש, וכרגע: ברמת מובהקות מתקבלים או שדוחים.

10 הרצאה 12: רגרסיה מתקדמת

10.1 רגרסיה לא ליניארית

במקרה שבו למשל יש לנו ערכי x_i, x_i^2, y נקבל כי $y = w_1 x + w_2 x^2 + b$. נשערך בהתאם באמצעות המטיריצה גם את w_1 וגם את w_2 .

10.2 רגרסיה של סדר ($Rank$)

ברגרסיה ליניארית רגילה אנחנו מניחים כי הקשר קו ישר והנתונים מתפלגים נורמלית, אך מה אם יש ערכים חריגיים או שהם אינם מתפלגים נורמלית? מקרה זה הוא מקרה שם חשוב רק הסדר של הנתודות במדגם, ולא הערך שלהם (או אם יש ערכים שמספרים את הנחת הנורמליות). ככלומר, אם למשל הנתונים הגיעו לנו לפי מקומותם בתחרות: דמי הגיע ראשון, נעם הגיע שנייע וכו' הלאה. במקום להריץ רגרסיה על המספרים המקוריים ((x_i, y_i)) אנחנו הופכים כל מספר לדרגה שלו במדגם:

1. מדרגים את כל ערכי X מזקZNן לגדול.
2. מדרגים את כל ערכי Y מזקZNן לגדול.
3. מחשבים לכל תצפית את ($d_i = rank(x_i) - rank(y_i)$)
4. מחשבים: באשר n הוא מס' הנתודות במדגם

$$r_{rank} = 1 - \frac{6 \sum d_i^2}{n(n-1)}$$

r הנ"ל שקול למתרם פרסון. הוא מתאים לפי סדר.

10.3 תיקון R^2 לגודל המדגם

ככל שנוסף יותר משתנים ב"ת למודל, הוא יטה להתאים יותר לנ נתונים ולכן R^2 גדל. (ככל שיחיו יותר משתנים, יהיה יותר רעש, אנחנו יותר נקרב למורות שזה לא באמת. לכן יש לנרמל את R^2 בנגד העובדה שמוסיפים עוד משתנים ב"ת.)
לכן נתקן את R^2 . נקבל כי:

$$\bar{R}^2 = 1 - (1 - R^2) \times \frac{n - 1}{n - p - 1}$$

באשר n מס' הנקודות במדגם ו p הוא מס' המשתנים הב"ת.
מהי המשמעות של $\bar{R}^2 \leq R^2$? נבחן כי הוא יכול להיות קטן מפיס עקרונית, וכן תמיד $\bar{R}^2 \leq R^2$ בהכרח. \bar{R}^2 עליה רק אם העליה ב R^2 גודלה מזו הצפואה באופן אקראי. כלומר: אם המשתנה החדש סתם אקראי \bar{R}^2 ירד ואמ' הוא באמת מסביר שהוא \bar{R}^2 עליה.

Identifiability 10.4

אם $X \in \mathbb{R}^{n \times p}$, עד כה הוכיחנו כי $p = \text{rank}(X)$ (לכן ניתן היה לחשב את $(X^T X)^{-1}$ - שכן $X^T X$ מסדר $p \times p$. בשיביל שהיא תהיה הפיכה $= \text{rank}(X)$. אם כן, אם $n > p$, ככלומר ישנים יותר משתנים מותצפים, $n \leq p$ ולכן היא בהכרח לא תהיה (p) מה קורה כאשר $n > p$? במקרה זה:

1. w היא *not Identifiability*
2. אפשר לבצע חיזוי (יש פתרון לבני האננסים הנוטנים), אך המודל חסר ערך (כי יש אנסות פתרונות אחרים).

מדוע $n > p$? כמה אפשרויות יכולנו:

1. קורלציה בין המשתנים.
 2. משתנים זהים מופיעים במודל בשמות שונים. (טמפרטורה במעלות ובפרנהייט).
 3. יש פחות נתונים מאשר משתנים ב"ת. (למשל אם לוחמים בדיקה גנטית, יש 4 מיליון עמודות - מקטיעים של DNA ומשם השורות הוא האנשים עליהם יש לנו את הדגימה - מדובר על כמה מאות. כמובן שם' המשתנים גדול מאוד ובמקרה זה $n > p$).
- במצב זה לא נוכל להחליט איזה מהמשתנים בקורסציה עם המשתנה התלוי.

אם נרצה כמה המצביע חמור (מבחינת *identifiability*) אפשר להשתמש ב- *VIF* .
Variance Inflation Factor *collinearity*

Variance Inflation Factor 10.5

כשנרצה לבדוק כמה המשתנים שלנו מדברים זה עם זה, במקומות לדבר עם המשתנה התלוי y , נשתמש *VIF* .
כלומר: כאשר השונות של מקדם הרגרסיה מתנפח בגלל מתאם גבוה בין שני משתנים מסוים. **מכיון הבעה מגיעה? בשיש מתאם גבוה בין שני משתנים, קשה למודל לדעת להחליט למי מהם לחת את הקזריט על השינוי ב- y .**

נניח כי אנחנו מחשבים מודל רגרסיה מהצורה:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

מבצע את השלבים הבאים:

1. נבנה n מודלי רגרסיה, כאשר בכל פעם אחד המשתנים יהיה התלויה. לדוגמה: $x_1 = \alpha_0 + \alpha_2 x_2 + \dots + \alpha_n x_n$
2. נסמן ב- R_i^2 את R^2 של המודל i .
3. נחשב לכל i את $VIF_i = \frac{1}{1-R_i^2}$

ערכים מקובלים ל- VIF :
 $collinearity$ גורר שאין בעיה $VIF = 1$
 יש בעיה $VIF > 1$
 יש בעיה משמעותית. $VIF > 10$

האינטואיציה: אם $R_i = 0$ כלומר לא הצליח להסביר אותו באמצעות האחרים, זה המצביע הטוב לנו, בו $1 \leq VIF < 10$ ואין לנו בעיה.

Ridge Regression 10.6

אם $p < rank(X)$ אין אפשרות להכניס אילוץ על פונקציית המטריה של הרגרסיה כדי לגרום למטריצה $X^T X$ להיות ההפיכה. בגרסה זו של רגרסיה, נמזרע את השגיאה הריבועית תחת האילוץ c באשר c פרמטר. בעזרה כופלי לגראנז נכל כתוב את הפונקציה כ:

$$\min_W (y - XW)^T (y - XW) + \lambda(W^T W - c)$$

כאשר λ כופל לגראנז. אם נגזר נקבל:

$$W = (X^T X + \lambda I)^{-1} X^T y$$

למעשה כתה כיוון נוספת מטריצה ייחידה עם פקטור λ , היא כתה מטריצה ריבועית ההפיכה.

10.7 מה עושים אם חלק מהנתונות היסוד של רגרסיה לא מתקינות?

הנתונות היסוד של רגרסיה יינארית: חזויו למשתנים תלויים רציפים, וריאנס קבוע לכל רמה של המשתנים הבלתי תלויים (*homoscedasticity*), השגיאה מתפלגת נורמלית. בעולם האמתי לעיתים קוראים דברים אחרים: משתנה תלוי קטגוריאלי (חוליה או בריא?), משתנה תלוי של ספירה (כמה ביקורים אצל הרופא?), התפלגות לא סימטריות. הפתרון הוא *GLM*. מרחיבים את הרגרסיה הלינארית בשתי צורות:

1. פונקציה מושנית.
2. ההתפלגות של המשתנה התלויה היא משפחה אקספוננציאלית כלשהו.

לשליטה רכיבים:

1. המשתנה התלויה שמעירously מושפח אקספוננציאלית. כלומר, אפשר לבטא את פונקציית הפילוג אקספוננט. לדוגמה: נורמלית, פואסון, ביןומי
2. חזויי ינארית: $\eta = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$
3. פונקציה מושנית $\eta = g(\mu)$

מדובר צורך פונקציה מושנית? אם המשתנה התלו依 אינו רציף, לא נוכל לבנות מודל באופן 'ישיר' $Y = X\beta$.

דוגמא לפונקציה מושנית: עבור רגרסיה יינארית הפונקציה הינה פונקציית היחידה $\mu = g(\mu)$. עבור רגרסיה לוגיסטיבית שבה $y \in (0, 1)$ נבחר $y = \log(\frac{1}{1-\mu}) = \log(\frac{1}{1-\mu}) \cdot g(\mu)$. עבור רגרסיה פואסיאנית בה $y \in (0, \infty)$ נבחר $y = \log(\mu) = g(\mu)$.

11 הרצאה 13: AB TESTING

כיצד עושים בדיקת השערות בתעשייה באמת? על זה נעה בהרצאה זו.
מה זה *AB TESTING*? בכל פעם שאנו מוחפשים בגול, בכל פעם שאנו נכנסים לשירות
כלשהו של גול אנחנו משתמשים בכמה שירותים/APIים ניסויים בו זמינות מוביל שאנו יודעים זאת. אלו
ניסויים שמתבצעים בעולם האמיתי. תעשיית הניסויים חיה ומוטבסט על כך שכן לבעז ניסויים
בצורה זו. מה האלטרנטיבה? להקשיב לדעתו של מישחו אחד: ובכן - אל: עליינו לנשות ולראות
נתונים לפני שנבצע החלטות.

11.1 הגדרה

הגדרה: ניסוי *AB* הוא ניסוי שנעשה אונליין (לא במעבדה) שנעשה אקראית (מגרילים אנשים לאחד
משני טיפולים לפחות), נתונים לניסוי לרוץ ובודקים מהם יקרה לבסוף.
במייקרוסופט עושים לפחות 400 ניסויי *AB* ביום. בשנת 2008 נעשו בממוצע 20 ניסויים
ביום. נשים לב כי התפיסה השתנה: בעבר גישה *HIPPO's paid person in the room* תפסה:
Highest כיוון, מוצעים על כל פיצ'ר קטן ניסוי לבדיקה על משתמשים אקרים.

מדוע נרצה לעשות מבחן *AB*? אנחנו די גורועים בלחוזות מה לעבור, ולכן נרצה לעשות ניסויים
שבילם לקבל החלטות מבוססות על נתונים ולא על דעתו של האנשים השימושים בחדר.
בהרבה מקרים זה אפשרי וזה הדרך היחידה לקבל החלטות מבוססות על נתונים. בתעשייה מדעי
המחשב - קל יותר לבעז ניסויים כאלו אונליין בניגוד לניסויים של רפואה.

כיצד ממבצע ניסוי *AB*? יש לנו אוכלוסייה (גודלה תלוי אם זו חברה או סטראטיפ) ואנו מגרילים
באופן אקרייל חלק מהמשתמשים לטיפול החדש שלנו (נניח, שינוי צבע הפונט בגול), נסמן טיפול זה
ב*B*. שאר המשתמשים: יקבלו את ניסוי *A* (המצב הנוכחיים). נרץ את שני הניסויים במקביל, ונבדוק
מה קרה למשתמשים שקיבלו את ניסוי *A* ומה קרה למי שקיבלו את ניסוי *B*.

לפני שMRIIZIM ניסוי *AB* אנחנו ממבצעים בדיקה למערכת שאכן ניתנת לביצוע. הניסוי יקרה
AA - TEST, ומטרתו היא לבדוק את מערכת הניסוי, לחשב את השונות בסביבת הניסוי ולהשוו
רוויזיון. (הניסוי זהה רק שהפעם כולם יקבלו את ניסוי *A*). אנחנו מצפים כי משתתי האוכלוסייה
לקבל את אותן התוצאות (בערך, יתכן שיש רעש, נמדד את השונות) שכן דגמוני באקרייל אנשים
ונתנו להם את אותו הניסוי *A*.

מה אנחנו מעוניינים לבדוק? במקרה פשוט ביותר, נרצה לבדוק אם ישנו שינוימשמעותי סטטיסטי
בין הממצאים: $E(B) = \bar{x}_B - \bar{x}_A$.
לפי מה נבעז את הרנדומיזציה? לפי מי משתמש? לפי אדם? לפי *cookie*? לא תמיד זה אפשרי לבצע
לפי אדם (איך נדע את כל האנשים? יש לנו את *the user* בלבד), אך אפשר לבצע באמצעות *cookie*
או באמצעות שם המשתמש.
דבר נוסף שנרצה לשים לב אלוי - האם אנחנו יכולים לעקוב לאורך זמן הניסוי? אם למשל נעשה
לפי יתכן שהאנשים ימחקו את *cookie* שלהם לאחר ימים וזמן הניסוי שלנו יוגבל ליוםיים.
עלינו לדאוג שנוכל לעקוב על הניסוי לאורך זמן.

סיכום בחירות גודל המדגם:

הדבר הכי יקר עבור חברות שרצו להריץ ניסוי הוא גודל המדגם. המשאבות שכולם מותחרים עליו
הוא כמה משתתפים בניסוי. יש לבחור פשרה בין גודל המדגם לבין אורך זמן הניסוי.
הניסוי יהיה חייב לרוץ מספיק זמן בשבייל להפחית את *novelty effect* (אפקט החדשנות) -
כאשר אדם נתקל בפיצ'ר חדש, ליקח לו זמן להתרגל אליו, מכמה כיוונים: ראשית, להכיר את המשחק

חדש, ושנית, אנשים בהתחלת יתלהבו מהמשחק החדש, ויתכו שבמהלך ההתלהבות תרד. עליינו גם להפחית השפעות של סופי שבוע או שעת ביממה - הניסוי צריך לróż מספיק זמן שנדיع כיצד המערכת מתנהגת בחגים וסופשים וגם באמצע השבוע.

כל שיש לנו יותר משתמשים, גבולות הסמק שלנו מוגבלים. אנחנו נרצה לומר כי התוצאה החדשה מוגבלת סטטיסטית לעומת תוצאות אחרות. לכן, בשביל להגעה לרווח סמק של 95% אנו זוקקים $\frac{16 \times \sigma^2}{\Delta^2} = n$ כאשר את σ^2 אנחנו יודעים (חישבנו בAA טסט) ואת Δ^2 (גודל האפקט - השני המינימלי שנרצה להיות) גם כן אנחנו יודעים.

אם נתחשב בכל הדברים הללו, בהרבה מהפעמים קיבל שאנו זוקקים מוגלים בגודל הרובה יותר גדול ממה שאנו יולים להרשות עצמנו (מבחינת משאבים) ולכן במקרה זה מדובר על **הגדלה הדרגתית**.

הגדלה הדרגתית (Ramping): מתחילה את הניסוי באחיזה נמוך של התעבורות, ואם הוא מצליח נלכ אל המנהלים שלנו, ואז נבקש 2%, אם זה גם צילוח, נעבור ל 5% וכן הלאה. מטרותיה של *Ramping* היא הזרמת סיכונים ומדידת הצלחה בlij להשתמש בתקציב גדול מדי. כמובן, שאם לאורך הזמן 5% הניסוי לא הצליח אנחנו נפסיק את ההגדלה. (הreasון הוא שאם לא הצלחנו ב 1% בודאות אין לנו סיכוי אבל אם הצלחנו ב 1% יתכן שלא דגמוני את כל האוכלוסייה, ולכן אנחנו צריכים להגדיל את הניסוי בשביל לקבל מוגלים מספיק גודל שבו גם קבוצות המיעוט נמצאות).

תנונו הניסיוני:

- * עליינו לאסוף כמה שיותר מדידות שונות, כדי לקבל נקודות מבט שונות.
- * עליינו להגדיר מدد אחד מראש שהוא שמעניין אותנו באמתו, נקרא לו *OEC*, והוא מدد ההצלחה שלנו, הוא יהיה אורך טווח. אם נבחר יותר ממדד *OEC* אחד, גם יש לזכור לתקן מדידות רבותות וגם הרבה יותר קל לנו לرمות את עצמנו.
- * דוגמאות ל*OEC*: במייקרוסופט *OEC* הוא מספר *Sessions* פר משותם. בנטפליקס - המدد שלהם הוא אחיזה הנטישה שלהם: כמה אנשים בחודש (בחוז) מתקשרים לבטל את המינוי שלהם. מדובר במידדים טובים - קשה לשנות אותם בטוח ורחוק. (רוב החברות לא מפרסמות את המיד OEC שלהם). כאשר אנחנו מפרסמים מראש את *OEC* היסכוי שלנו לرمות את עצמנו ו/או הציבור קטן מאוד.

11.1.1 סוגים מדדיים (סוגי *OEC*)

- א. ממד ייעדים - ממד שמעניין את הארגון והוא מאוד חשוב לו.
- ב. ממד ביןים - מדדים קצרי טווח שאנו חונכו חשובים שמנגנים את הממד שמעניין אותנו באמתו.
- ג. ממדים אילוציים - מדדים לפועלות שלא נרצה שייקרו.

11.1.2 קритריונים למדדיים

- א. נרצה כי יהיה קל ופושט למدوוד את הממד
- ב. נרצה שם יהיו יציבים עם סטיית תקן נמוכה.
- ג. נרצה מדדים שקשורים ליעד
- ד. נרצה מדדים שכן ניתנים להשפעה במידה מסוימת (למרות שהם יציבים)
- ה. נרצה שהם יהיו מספיק רגילים (אם הפיצ'ר שהכננו עשה שינוי במדד שבדקנו יוכל לראות זאת)
- ו. חסינים למנהגולציות (זו הסיבה שרוב החברות לא יפרנסו את *OEC* שלהם. מה הבעה? אנחנו יודעים כתה מה שוכواب לנטפליקס - ולכן אנחנו יבואו לנציג השירות של נטפליקס ונגידו להם אני רוצה הנחה או שאעוזב. זו בעיה עברו נטפליקס).

חוק Goodhart's: כשמכרים על הממד בתור המטרה, הוא אינו ממד טוב עוד.

11.1.3 השפעת הניסוי על המשתתפים

- אפקט הראשוניות:** כמו זמן לוקח למשתמשים להתרגל לממשק החדש. עשוי לעזור לקבוצת הביקורת.
- אפקט החידוש:** משתמשים אוחבים לנסות דברים חדשים. עשוי לעזור לקבוצת הניסוי.
- אפקט העברה:** שאריות ההשפעה של הממשק הקודם בזמן שהמשמש החדש החל לפעול (עד שימושם מוגרלים לממשק החדש).

11.1.4 הרצת ניסויים במקביל

- אם הניסוי כל כך יקר, אולי שנרץ ניסויים במקביל? ובכן זה מה שקרה בזמן בוגר. בכל רגע נתנו המשתמשים נמצאים בכמה מאות ניסויים. כיצד נדע שהניסויים לא ישפיעו זה על זה?
- א. בחירת מבחןנים שפועלים על איזוריים שונים בשירות (נגידיל את אותן משתמשים לניסויים שלא קשורים זה לזה)
- ב. נחלק את המשתמשים לקבוצות זרות זו לזו
- ג. נוכל לבדוק את ההתנגשות על אוכלוסייה קטנה

נרצה להסתכל לפעם על תתי אוכלוסיות. למשל תתי אוכלוסיות שיעניינו אותנו: מדינה או שוק של קונים, מכשיר (כמו נייד או טבלט), זמן בשבועו או יום וסוג משתמש.

עלינו להזהר **מפרזוקס סימפסון**: יתכן שאם נסתכל על כל האוכלוסייה נראה אפקט בכיוון אחד (למשל קו עלייה) אך אם נסתכל על תתי אוכלוסייה נראה אפקט אחר - של ירידה. זה חורע על עצמו בצעירים מול מבוגרים, שمعدיפים דברים שונים.

11.1.5 טעויות נפוצות בניסויי AB

- א. הטעות העיקרית היא שלא בחרנו מספיק משתמשים והאוכלוסייה שלנו קטנה מדי וגבולות הסטם לא מספיק מופרדים בשביב לקבל החלטה.
- ב. ה策זה לתוצאות הניסוי לפני סיום הניסוי (למשל להפסיק את הניסוי לאחר חצי מיליון משתמשים, قبل הכספי, אם התוצאה מספיק מובהקת סטטיסטיית אז עצור את הניסוי). מעשה במקומות להסתכל רק בסוף - הסתכנו בامي, וכך רימינו את עצמנו. (למה? כי חילקנו את $P-value$ במליאן ולא בחצי מיליון).
- ג. מנהיגים שהעולם יציב מדי. למשל: לא עשינו מבחן AA או שעשינו לפני שנה ואנחנו מנהיגים שהעולם יותר כשהיה.
- ד. הטיתת הרזרות: נניח ואנחנו מכינים פיצ'ר חדש, שהאנשים לא סובלים אותו. כולם נטושו אותו, פרט לבודדים. מי כן ישר? ה-10% שכן אהבו אותו.
- ה. לפעמים קשה לעקוב אחרי ייחדות המדידה.
- ו. יתכן שיש השפעות ייצוגיות (למשל באיראן למשתמשים אין אינטרנט כרגע, ונניח ונדגום את אוכלוסיית איראן, מי משתמש כרגע לא אוכלוסייה מייצגת).
- ז. חifyבים לישם את הטיפול החדש - علينا ממש לקודד את הטיפול החדש וזה דבר יקר מאוד.
- ח. עקביות בחווית המשתמש - נרצה שלמשתמש תהיה חוות עקבית, כאשר אנחנו עושים הרבה ניסויים במקביל למשתמש אנחנו פוגעים בחווית המשתמש.