שיטות סטטיסטיות ־ סיכום הרצאות למבחן

13 בינואר 2026

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

תוכן עניינים

1 הרצאה 1: מבוא לקורס

סטטיסטיקה: תחום ידע שנוגע לאיסוף, עיבוד ניתוח והסקת מסקנות מנתונים כמותיים. מחלקים את הסטטיסטיקה לשני תחומי דעת: תאורית, והיסקית.

סטטיסטיקה תאורית: עוסקת בתיאור תמציתי וקל לתפיסה של אוכלוסייה על סמך מדדים. למשל: ייצוג ע"י דיאגרמה, מדדי מיקום כמו ממוצע שכיח וחציון, מדדי פיזור כמו שונות וסטיית תקן.

סטטיסטיקה היסקית: עוסקת בנסיון להגיע למסקנות לגבי אוכלוסיה על סמך מדגם. (למשל: סקר בחירות)

אמידה סטטיסטית: אלו שיטות מתמטיות שמאפשרות לגזור מתוך נתוני המדגם אומדן ערך של משתנה עבור אוכלוסיה. הבסיס ללמידה חישובית.

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

אמפירי = מבוסס על ניסוי

1.1 שיטות מחקר:

כיצד אנו רוכשים ידע על העולם?

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

הגישה האמפירית: ידע מבוסס על תצפית נסיון ומדידה. (למשל: השמש זרחה הבוקר, היא תזרח גם מחר).

הגישה המדעית = הגישה האמפירית + הגישה הרציונלית.

מטרת הגישה המדעית: להבין עבר, לנבא עתיד ובעיקר לנסח תאוריות.

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

ניסוח בעיית מחקר: בעיית מחקר היא בעיה שניתן לחקור אותה בכלים מדעיים. הבעיה צריכה

להיות מנוסחת בצורה אובייקטיבית, ברורה וחד משמעית. הבעיה צריכה לבטא יחס בין שניים או יותר משתנים. הבעיה חייבת לעמוד ב**בחינה אמפירית** (דרך למדידת משתנים)

השערת מחקר - ממוקדת 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) $\Leftarrow$ תיאור והצגה (הרצאה 3) $\Leftarrow$ אומדן פרמטרים (הרצאה $4-6$) $\Leftarrow$ בדיקת השערות (הרצאה $7-9$) $\Leftarrow$ ניסוח השערה $\Leftarrow$)וחוזר על עצמו)

אומדן מידע + תיאור והצגה = סטטיסטיקה תאורית.

אומדן פרמטרים + בדיקת השערות = סטטיסטיקה היסקית.

2 הרצאה 2: סטטיסטיקה תאורית 1

2.1 איסוף מידע

2.1.1 ממי אוספים את המידע?

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

ב. **מדגם** - תת קבוצה (מייצגת) של האוכלוסיה

1. אם קיים קושי במדידה של האוכלוסיה כולה (מסובכת, ארוכה, יקרה)
2. קושי באיסוף המידע (הרבה מידע)
3. עצם המדידה פוגע בתכונה (כלומר, אם למשל אנחנו במפעל גפרורים ורוצים לבדוק את מס' הגפרורים התקינים - בשביל לבדוק אם הוא תקין נצטרך להשתמש בו ולכן הפכנו אותו ללא תקין. אם נקח את כל הגפרורים ונבדוק אותם נשאר ללא גפרורים תקינים: לכן אנחנו חייבים לקחת מדגם).

מה הכוונה ב**תת קבוצה מייצגת?** קבוצה שמשמרת את התכונות של האוכלוסיה, משמרת את הפיזור וניתן להכליל ממנה.

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

2.1.2 מה אנחנו אוספים?

משתנה: תכונה הניתנת לתצפית ומדידה עבור כל אלמנט באוכלוסיה.

ערך: הערך שנמדד עבור אלמנט יחיד באוכלוסיה.

מידע: הערכים שנמדדו עבור כל האוכלוסיה.

2.1.3 לשם מה אנחנו אוספים את המידע?

סטטיסטי: ערך המחושב על סמך הדאטא, כלומר על סך כל הערכים שנמדדו. (ממוצע הדגימות).

פרמטר: מאפיין של האוכלוסיה. למשל, תוחלת ההתפלגות.

—♡- המשתנה הוא תכונה, למשל אם נבצע מדגם אודות סכום הכסף הממוצע שסטודנט מוציא בשנה א', וקיבלנו שהממוצע הוא $178\$$, אזי הסטטיסטי הוא $178\$$ וכן המשתנה הוא סכום הכסף הממוצע.

—♡- **יתכן סטטיסטים שונים:** למשל מינימום מקסימום, חציון, שונות וכו'.

2.1.4 שיטות דגימה הסתברותיות

בשיטות דגימה הסתברותיות ישנה הסתברות שווה לכל פרט להבחר.

1. דגימה אקראית ־ רנדומית. דגימה של k איברים מתוך N. זו דגימה שיכולה להתבצע עם החזרה או ללא החזרה. **בקורס זה כאשר נאמר כי אנו מודדים ־ נמדוד לפי דגימה אקראית.**

2. דגימה בשכבות ־ חלוקת האוכלוסיה לשכבות זרות ומשלימות. דגימה רנדומית (לפי פורפורציה) מכל שכבה. דוגמה לדגימה בשכבות: סקרי בחירות. לוקחים שכבות אוכלוסיה ־ אם יודעים שישנם 27% מהאוכלוסיה בגילאים $50-40$ אזי דוגמים פורפורציונלית משכבת גיל זו.

3. דגימת אשכולות ־ חלוקת כל האוכלוסיה לקבוצות זרות ומשלימות. דגימה רנדומית של קבוצות והוספת כל הפרטים מכל קבוצה למדגם. למשל: ביצוע סקר מדד האושר. במקום למדוד אחד אחד, אפשר למדוד בתי אב. אם בית אב יצא כ$5/5$ במדד האושר ־ כל האנשים בבית האב הנ"ל ייחשבו כ$5/5$ במדד האושר. כלומר ־ לוקחים את כולם.

2.1.5 שיטות דגימה לא הסתברותיות

ישנן שיטות דגימה שאינן הסתברותיות.

1. **דגימת נוחות ־** "מן המוכן", הכל בבת אחת. כלומר ־ מקבלים את הדגימה בבת אחת. למשל: משאל רחוב, מקבלים את התוצאות מיד. מה טוב בשיטה? מהיר. מה בעייתי? לא מייצג את האוכלוסיה.

2. **דגימה שיפוטית ־** לפי שיקול דעת החוקרת, לפי מענה על שאלונים. מה טוב בשיטה זו? אנחנו מניחים שהחוקרת יודעת מה היא עושה ולכן זה טוב לנו שהיא בוחרת את האוכלוסייה. מה לא טוב? לא מייצג וסיכוי גבוה להטייה.

3. **דגימת כדור שלג ־** "חבר מביא חבר". כלומר ־ ניסויים שאדם מגיע, מקבל כסף על הניסוי ואומרים לו להביא חברים לניסוי שיבוא גם הוא "להרוויח כסף". יתרון: קל ומהיר, דגימת אוכלוסיה זהה. חסרון: לא מייצג, ישנה הטיה, מדגם של חלק ספציפי באוכלוסיה.

ישנן סכנות לתקפות הניסוי: דגימה לא מייצגת/ מוטה, דגימה "התנדבותית", דגימה קטנה מדי.
למה להשתמש בשיטות דגימה לא הסתברותיות? פיילוט, מיעוט אקוטי, תופעות מאוד נדירות.

בקורס זה נשתמש בשיטות דגימה הסתברותיות.

2.2 משתנים

2.2.1 סוגי משתנים

א. קטגורי: קבוצת ערכים סופית. קטגורה מדגרית, דרגה בצבא, קבוצת המידות=$\{XS, S, M, L, XL, XXL\}$.

ב. מספרי: מס' הסטודנטים בקורס, מספר אסיסטים למשחק, גובה משקל וכו'.
המשתנה הבדיד: קבוצת ערכים סופית ובת מנייה.
המשתנה הרציף: קבוצת ערכים אינסופית, בין שני ערכים קיים ערך. למשל ־ מרחק.

2.2.2 סיווג משתנים ־ סולמות מדידה

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

סולם סדר: יחס זהות, עם יחס סדר. למשל: תווית מידה, דרגה אקדמית. ־ משתנה קטגורי. בסולם זה כן יש יחס זהות, כל אחד משתייך לדבר מסויים אך יש יחס סדר בין הדברים.

סולם רווחים: עם יחס סדר, עם מרווחים קבועים. למשל: טמפרטורה ־ משתנה מספרי. בסולם זה: יש משמעות למרווחים בין הערכים. למשל בטמפרטורה יש משמעות למרווחים בין הטמפרטורות השונות.

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

2.2.3 מדדים סטטיסטים

סטטיסטי: ערך המחושב על סמך התצפיות בפועל. למשל - ממוצע, חציון.
פרמטר: תכונה של האוכלוסיה המקורית. למשל - תוחלת בהתפלגות נורמלית, פורפורציה בהתפלגות בינומית.
♡- בחלק של "סטטיסטיקה תאורית", נשתמש בסטטיסטים לתיאור הדאטא. בחלק "הסקה סטטיסטית" נשתמש בסטטיסטים כאומדן לפרמטרים.

2.3 תיאור והצגה

כיצד ניתן להציג את המידע שנאסף?
$*$ תצוגה טבלאית

1. טבלת שכיחויות. למשל פונקציה $f : \{0,1,...,100\} \to \mathbb{N}$ שמקבלת ציון u ו$f(u)$ זה מס' הסטודנטים שקיבלו אותו.
2. שכיחות יחסית: rf. בטבלה מטה סה"כ שכיחות שמסתכמת ל20. שכיחות יחסית תהיה האחוז של הערך כ

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

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

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

4. **משתנה מספרי בדיד: חלוקה למחלקות**

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

5. **משתנה מספרי רציף: חלוקה למחלקות**

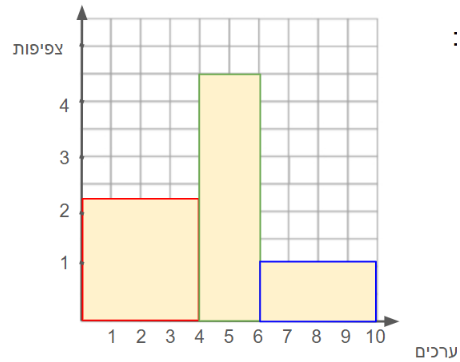
הגבול העליון של מחלקה אחת מתלכד עם הגבול התחתון של זו שאחריה. בהינתן מחלקה $[x_0, x_k]$
רוחב מחלקה: ההפרש בין גבול עליון אמיתי לגבול תחתון אמיתי. יתקיים $I = x_k - x_0$
מחלקה פתוחה: רק גבול עליון או תחתון

מחלקה	שכיחות f	רוחב מחלקה I	מרכז טווח	צפיפות d=f/I
0-4	9	4-0 = 4	0 + 4/2 = 2	9/4 = 2.25
4-6	9	6-4 = 2	4 + 2/2 = 5	9/2 = 4.5
6-10	4	10-6 = 4	6 + 4/2 = 8	4/4 = 1.0

מרכז טווח של מחלקה $[x_0, x_k]$**:** $x_0 + \frac{I}{2}$
צפיפות המחלקה מוגדרת להיות: $d = \frac{f}{I}$

מכאן מקבלים **היסטוגרמה**: גרף שמייצג את הערכים. השטח מתחת להיסטוגרמה הוא סה"כ השכיחויות.

:



כיצד בונים היסטוגרמה?

א. מחליטים על מס' המחלקות שנרצה -k
ב. מחשבים את הטווח של ההיסטוגרמה $r = max - min + 2$ (מוסיפים פלוס 2 רק באשר אנחנו יודעים את הערכים עצמם ממש ולא היסטוגרמה).
ג. מחשבים רוחב כל מחלקה $a = \frac{r}{k}$
ד. מחשבים גבולות מדומים - $min - a$
ה. בחירת יחידת הדיוק u
ו. חישוב גבולות אמיתיים $min - u$

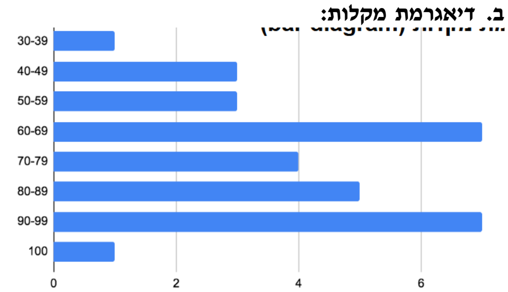
היסטוגרמה נכונה רק כאשר נתונים לנו כל הנתונים. אם נתון לנו טבלת שכיחויות אי אפשר לעשות זאת.

* תצוגה גרפית:
מייצגים נתונים באמצעות דיאגרמה.

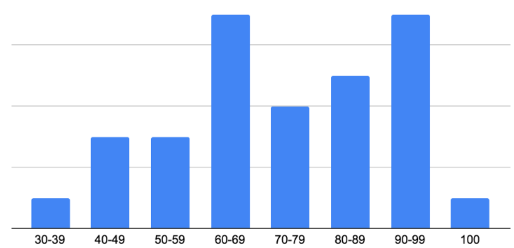
א. **דיאגרמת גבעול־עלה:** דיאגרמה בה מפצלים את הערכים לעשרות ויחידות. חלוקת טווח הערכים לגבעולים. וכן פירוט ערכי העלים.

Stem (ספרת 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 4 6
10	0

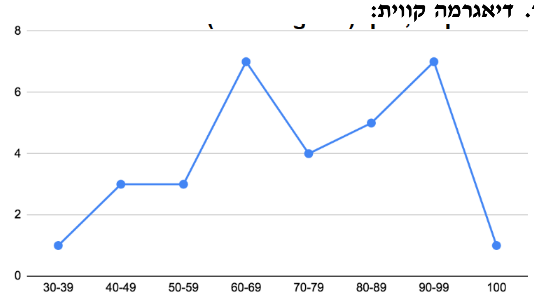
ב. דיאגרמת מקלות:



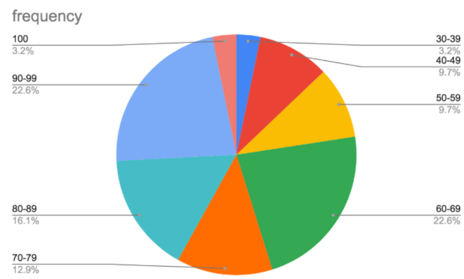
ג. דיאגרמת עמודות:



ד. דיאגרמה קווית:



ה. דיאגרמת עוגה:



מתי נשתמש באיזו דיאגרמה?

עבור כל סולמות המדידה: טבלת שכיחויות, דיאגרמת עמודות, תרשים עוגה.

עבור סולם רווחים או סולם מנה: היסטוגרמה, דיאגרמת גבעול עלה, דיאגרמת קופסא.

$*$ **מדדים סטטיסטיים:**

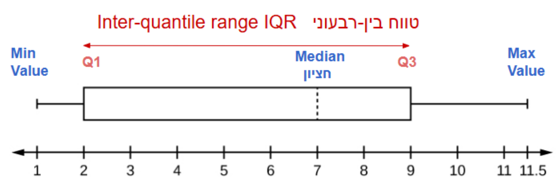
שכיח: הערך עם השכיחות הגבוהה ביותר.

מרכז הטווח: הממוצע בין התצפית הגבוהה ביותר והנמוכה ביותר.

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

ממוצע: סכום כל הערכים מחולק במספר התצפיות.

דיאגרמת קופסא:
בשביל לחשבה מסתכלים על ערך המינימום, המקסימום, החציון וכן Q_1 שיהיה החציון של החצי הנמוך (מהמינימום עד החציון) וQ_2 שיהיה החציון של החצי הגבוה (מהחציון אל המקסימום).



כיצד מציירים דיאגרמת קופסה?

א. מסדרים את הנתונים לפי סדר עולה

ב. מוצאים מינימום, מקסימום, חציון , רבעון ראשון ורבעון שלישי

ג. מציירים לפי הנתונים שמצאנו קודם – בין הרבעון הראשון לרבעון השלישי אנחנו מציירים קופסה, בתוכה מסמנים את החציון. הטווח שבין הרבעון הראשון לרבעון השני נקרא טווח בין רבעוני

ד. לאחר מכן מציירים קווים לקצה הטווח – בין הרבעון הראשון למינימום ובין הרבעון השלישי למקסימום

הערה: למציאת החציון - אם יש לנו מס' אי זוגי זה קל, האמצעי. אם יש מס' זוגי יש שני חציונים - החציון בקורס יוגדר להיות ממוצע שני החציונים הנ"ל.

3 הרצאה 3: סטטיסטיקה תאורית 2

3.1 מדדים סטטיסטיים

סטטיסטי הסדר: יהיו $X_1, ..., X_n$ משתנים מקריים עבור אוכלוסיה או מדגם. יהיו $x_1, ..., x_n$ הערכים שנמדדו עבורם בהתאמה. נסדר את הערכים $x_1, ..., x_n$ בהתאמה בסדר עולה. נקבל את סטטיסטי

הסדר:

$$x_{(1)}, ..., x_{(n)}$$

שכיח: הערך עם השכיחות הגבוהה ביותר,

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

מרכז הטווח: הממוצע בין התצפית הנמוכה ביותר לגבוהה ביותר.

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

חציון:

$$\overline{x} = \left\{ \begin{array}{cc} x_{(k+1)} & n = 2k+1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)} & n = 2k \end{array} \right\}$$

ממוצע: סכום הערכים חלקי מס' התצפיות

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

3.2 חישוב מדדי מיקום מרכזי עבור מחלקות עם גבולות אמיתיים

כאשר נתונה לנו טבלת שכיחויות וגבולות אמיתיים, נסמן $f(v)$ כשכיחות של v, את השכיחות היחסית ב$rf(v)$ ואת היחסית המצטברת $RF(v)$.
נחשב את הממוצע כך:

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

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

כיצד נחשב את החציון בטבלת מחלקות רגילה? נסתכל על השכיחות היחסית המצטברת $RF(v)$, ונחפש היכן אנחנו פחות מחצי מהקלט, והחציון יהיה שורה אחת אחריו. אם קיים ערך עבורו $RF(v) = 0.5$ אזי החציון יהיה הממוצע של זה לפניו וזה אחריו.
וכיצד נחשב חציון עבור מחלקות עם גבולות אמיתיים?
מחלקה m, גבולות $L_0 - L_1$, שכיחות f ומצטברת F.

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

הרעיון יהיה למצוא את האיבר אשר הקו מתחתיו מחלק את ההיסטוגרמה לשני חלקים שווי שטח. בשלב הראשון נצטרך לזהות את המחלקה m בה החציון אמור להמצא, את זה נעשה כמו שעושים בטבלה רגילה. נסמן את הגבולות שלה ב$L_1 - L_0$. וכן n מס' התצפיות.

באופן דומה, לחשב את הרבעונים: נזהה את המחלקה m_1 בה נמצא הרבעון ונחשב – 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:

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

ואלפיון 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 אמיתיים (גבול עליון של מחלקה קודמת זהה לגבול תחתון של מחלקה נוכחית).

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

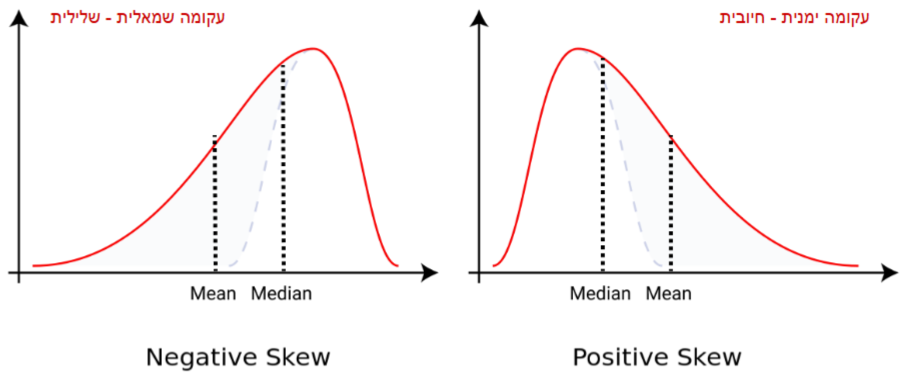
אם נבחר במדד מסויים, מהי פונקציית ההפסד שלי?
א. מס' השגיאות: כמה מהערכים אינם שווים למדד עצמו $|\{x_i | x_i \neq \overline{x}\}|$:
כאשר נסתכל על פונקציה זו, השכיח ימזער את מס' השגיאות. כלומר אם הפונקציית הפסד שמעניינת אותי היא מס' השגיאות אזי נשתמש בשכיח.
ב. השגיאה המקסימלית: המרחק המקסימלי מהמדד עצמו $max_i|x_i - \overline{x}|$
כאשר נסתכל על מדד זה, מרכז הטווח ממזער את השגיאה המקסימלית.
ג. סכום השגיאות המוחלטות: מרחקים אבסולוטיים של כל הערכים הממדד $\sum_i |x_i - \overline{x}|$
החציון ממזער את סכום השגיאות המוחלטות.
ד. סכום ריבועיי השגיאות: מרחקים ריבועיים של כל הערכים מהמדד $\sum_i (x_i - \overline{x})^2$
הממוצע מפחית למינימום את סכום ריבועי השגיאות.

מכאן נבין כי כל פונקציית הפסד מתייחסת ו"מענישה" מדד אחר. לכל שימוש ישנו מדד שונה שטוב עבור $\overline{x}$.

תכונות מדדים סטטיסטיים למיקום מרכזי $\overline{x}$:

	שכיח	אמצע טווח	חציון	ממוצע
פונקציית הפסד	מספר שגיאות	שגיאה מקסימלית	סכום השגיאות המוחלטות	סכום ריבועי השגיאות
רגישות לערכי קיצון	אין	רבה	מעטה	רבה
סולמות מדידה	שמי ומעלה	רווחים ומעלה	סדר ומעלה	רווחים ומעלה
שימושיות בהסקה	בינונית	פחותה	פחותה	מרובה

נשים לב. בעקומת פעמון סימטרית: הממוצע=חציון=שכיח.
בעקומת פעמון אי סימטרית שמאלית (הזנב לצד שמאל) : ממוצע < חציון < שכיח
ביקומת פעמון אי סימטרית ימנית (הזנב לצד ימין) : ממוצע > חציון > שכיח



3.3 מדדי פיזור

א. אחוז השגיאות: אחוז התצפיות השונות **מהשכיח** $\frac{1}{n}|\{i|x_i \neq \overline{x}\}|$

ב. גודל השגיאה המקסימלית: המרחק הגדול ביותר **ממרכז הטווח** $max_i|x_i - \overline{x}|$

ג. הטווח: המרחק בין ערכי קיצון

ד. הטווח הבין רבעוני: הטווח בו נמצאים 50% הערכים המרכזיים בהתפלגות. (מה שאנחנו מציירים בדיאגרמת Box).

ה. ממוצע הסטיות המוחלטות: ממוצע מרחקי התצפית **מהחציון.** $\frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|$

ו. ממוצע ריבועי הסטיות: ממוצע ריבועי מרחקי התצפית **מהממוצע** $\frac{1}{n}\sum_{i=1}^{n}|(x-\overline{x})^2|$. נשים לב כי העלאה בריבוע מענישה יותר את הקצוות, וזה יותר טוב עבור המרכז ולכן זה בודק היטב את הקצוות.

תכונות:

	אחוז השגיאות	**שגיאה מקסימלית**	**טווח**	**טווח בינרבעוני**	**ממוצע סטיות מוחלטות**	**ממוצע סטיות ריבועיות**
פונקצית הפסד	מספר שגיאות	שגיאה מקסימלית	–	–	סכום סטיות מוחלטות	סכום סטיות ריבועיות
מדד המרכז הנבחר	שכיח	מרכז טווח	–	–	חציון	ממוצע
רגישות לערכי קיצון	רגישות לכל הערכים	רגישות רק לערכי קיצון	רגישות רק לערכי קיצון	אין	יש	יש רגישות גבוהה
מהירות החישוב	מהיר	מהיר	מהיר	איטי	איטי	איטי
סולם המדידה	שמי ומעלה	רווחים ומעלה	רווחים ומעלה	רווחים ומעלה	רווחים ומעלה	רווחים ומעלה
שימושי להסקה	לא	לא	לא	לא	לא	כן

3.4 שונות וסטיית תקן

אנחנו נשתמש בעיקר בממוצע סטיות ריבועיות, הידועה בשמה: **שונות.**

עבור רשימת ערכים:

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

וסטיית התקן:

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

עבור טבלת שכיחויות:

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

יש לשים לב ־ השונות וסטיית התקין באוכלוסיה ובמדגם שונים זה מזה. באוכלוסיה, כפי שראינו למעלה.
במדגם:

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

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

מדוע מחלקים ב$n-1$ ולא בn? נגלה בהרצאה 5.

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

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

3.5 ממוצע משוקלל ושונות מצורפת

עבור k כיתות שונות, בהינתן N מס' התלמידים בשכבה מתקיים כי הממוצע המשוקלל הינו:

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

עבור k כיתות שונות, השונות המצורפת הינה:

$$S_T^2=\frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij}-\overline{x_T})^2=\sum_{j=1}^{k}\frac{n_j}{N}S_j^2+\sum_{j=1}^{k}\frac{n_j}{N}(\overline{x_j}-\overline{x_T})^2$$

החלק הימני בביטוי הוא השונות בין הקבוצות השונות, והחלק השמאלי היא השונות בתוך הקבוצות (סוכמים).

תיקנון: למשל, סטודנט קיבל 70 בחשבון ו75 בתנך. היכן הצליח יותר? בחשבון הממוצע היה 65 וסטיית תקן 3. בתנך 70 ו4 בהתאמה. כיצד נדע? ננרמל -

ציון התקן של x: מרחק מהממוצע הנמדד ביחידות סטיית התקן.

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

מכאן, נקבל שהתצפיות יהיו בתוך העקומת Z המפורסמת - התפלגות נורמלית סטנדרטית.

3.6 מדדי קשר בין מספר משתנים

למשל: האם יש קשר בין טמפרטורה ממוצעת באזור לתנובת עצי פרי באזור?
עד כה דנו במשתנה בודד, נדון כעת במס' משתנים.

יהיו לנו n תצפיות ובכל אחד מהתצפיות יש לנו ערכים (x,y). במערך ניסוי שכזה נרצה ללמוד על הימצאות הקשר בין x לy.

נוכל להשתמש בדיאגרמת פיזור: על ציר הx ערכי הx ועל ציר הy ערכי הy. לבדוק האם קיים קשר לינארי.

מקדם המתאם של פירסון: מדד קשר הממלא אחר הדרישות הבאות־
א. ערכו המוחלט יהיה מקסימלי באשר הקשר מושלם (כל הנקודות על הישר ־ הקשר לינארי)
ב. סימנו של המדד שלילי או חיובי יבטא את כיוון הקשר (חיובי כאשר חיובי ולהפך).

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

נזכר כי הגדרת השונות המשותפת:

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

ומכאן ש:

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

באשר אנחנו עובדים עם מדגם אנחנו נחלק ב$n-1$.

נרצה ליצור קו מגמה מהצורה $y = ax + b$. 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 וסולמות מדידה), אוספים נתונים (מאוכלוסיה ומדגם), מארגנים את הנתונים (הצגה טבלאית כמו שכיחות או צפיפות או ויזואלית כמו היסטוגרמה וכן מחשבים מדדים סטטיסטיים) ולבסוף מסיקים מסקנות. אם הנתונים נאספו על כלל האוכלוסיה אזי סיימנו. **אם הנתונים נאספו על מדגם מייצג מתוך כלל האוכלוסייה: עוד לא סיימנו.**

נרצה לשאול כמה שאלות חשובות. האם ניתן להכליל ממדדיים במדגם למדדים באוכלוסייה? באיזו רמת בטחון ניתן לבצע הכללה זו? האם לקבל או לדחות השערה ותחת אילו תנאים?

המטרה העיקרית בהרצאה זו תהיה לבדוק ־ האם יש קשר בין תופעות המדגם לתופעות באוכלוסייה? כיצד נעשה זאת: באמצעות הסתברות. נראה כיצד הסתברות וסטטיסטיקה נפגשים.

4.2 הסקה סטטיסטית

נזכר כי ישנה הגישה המדעית שמורכבת משתי גישות. האמפירית ("הכל מדיד"), והרציונלית: גישה שמבוססת על כללי היסק.

הסקה דדוקטיבית: (כלל $\Longleftarrow$ פרט), היסק לוגי, אמיתות ההנחות מחייבת את אמיתיות המסקנות. למשל: הנחה1- אין מים על כוכב הלכת חמה, הנחה2- ללא מים אין חיים. אז מסקנה: אין חיים על כוכב הלכת חמה. ניתן להפריך את ההנחות אך לא את המסקנה(!!)

הסקה אינדוקטיבית (פרט $\Longleftarrow$ כלל): 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 השמש זרחה כל בוקר, אז היא תזרח גם מחר. ניתן להפריך את המסקנה! (!!).

הבעיה המהותית: מה ההצדקה להסקה אינדוקטיבית במדע (כל המדע מתבסס על הסקה שכזו)? לא נלמד זאת בקורס - זה מדעי הדשא. עם זאת, הבעיה הכמותית: כיצד לכמת את מידת הוודאות שבתוך אי הוודאות? כן בקורס שלנו.

4.3 מושגים בסיסיים

משתנה מקרי: "תכונה" שהיא משתנה שלקוחה מהתפלגות מסויימת F. כלומר $X \sim F$.

תצפית: תוצאה של ניסוי מקרי מתוך המשתנה המקרי X.

דגימה: ביצוע רצף תצפיות (ניסויים) $X_1, ..., X_N$ באשר $\forall_{1\le i\le N} X_i \sim F$.

מדגם מקרי בגודל n מתוך מ"מ X: מדגם של n משתנים מקריים כך ש:

א. $X_1, ..., X_n$ הם מ"מ בלתי תלויים

ב. לכל מ"מ X_i יש את אותה פונקציית ההסתברות כמו של X , כלומר לכל i מתקיים $X_i \sim F$.

משפט: דגימה מקרית (אקראית, רנדומית) עם החזרה של n איברים מתוך אוכלוסייה עם תכונה $X \sim F$ שקולה למדגם מקרי בגודל n מתוך מ"מ מתאים $X \sim F$. ולהפך.

מסקנה: מבחינה מעשית ($\Longleftarrow$) נבצע דגימה מקרית מתוך אוכלוסייה גם כאשר בפועל נרצה לדגום ממ"מ. וכן מבחינה תאורטית ($\Longrightarrow$)נוכל להשתמש בכל מה שאנו יודעים על מ"מ על כל דגימה מקרית.

נשים לב: תכונה של האוכלוסיה נקראת פרמטר, וערכו קבוע אך לא בהכרח ידוע. מדד המבוסס על המדגם נקרא סטטיסטי וערכו ידוע אך לא בהכרח קבוע.

עבור אוכלוסיה בגודל k ניתן לייצר הרבה מדגמים שונים בגודל $n < k$ מכאן שכל סטטיסטי הוא משתנה מקרי עם התפלגות משל עצמו - ההתפלגות הזו נקראת ה**תפלגות הדגימה של הסטטיסטי**. (כלומר, תאסוף את כל המדגמים, כל אחד מהם מוציא סטטיסטי, כל סטטיסטי הוא משתנה מקרי עם התפלגות שלו.).

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

4.4 התפלגויות דגימה

צורת התפלגות הדגימה: צורת ההתפלגות תלויה במספר גורמים - בסוג ההתפלגות באוכלוסיה, בסוג הסטטיסטי ובגודל המדגם. לכן נקפיד כשנדבר על "התפלגות דגימה" : התפלגות הדגימה של סטטיסטי מסוים s עבור מדגמים בגודל n שנלקחו מאוכלוסיה בה ערכי המשתנה מתפלגים לפי התפלגות F.

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

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

הוכחה:

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

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

טענה: $Var[\overline{X}] = \frac{V[X]}{n}$ $(\sigma^2_{\overline{X}} = \frac{\sigma^2_X}{n})$
הוכחה:

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

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

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

מדוע זה מוצדק להשתמש במדגם אחד? ברור כי במדגמים שונים עבור אותה אוכלוסיה יש ממוצעים שונים, אם נדגום הרבה מדגמים ונחשב ממוצע לכל אחד, ממוצע הממוצעים יתקרב מאוד לממוצע באוכלוסיה. **אבל:** אין בכוונתנו לדגום הרבה מדגמים! אלא מדגם אחד ויחיד! אז: השאלה למעשה: מהי הסבירות שהממוצע במדגם שדגמנו סוטה (בהרבה) מהממוצע באוכלוסיה? שאלה שקולה: מהי הסבירות שהממוצע במדגם שדגמנו סוטה בהרבה מהתוחלת של הממ (ממוצע המדגם) עצמו? כלומר - כמה הערך שלי רחוק מהתוחלת של ממוצע המדגם. זו שאלה עדיפה לנו - כי כאן יש מדגם אחד בדיוק. לפיכך: נתעניין במידת הפיזור של התפלגות הדגימה של ממוצע המדגם. סטיית התקן של ממוצע המדגם שווה לסטיית התקן של הממ המקורי מחולקת בשורש n גודל מדגם ולכן: **ככל שהמדגם גדול יותר, שונות/סטיית התקן של ממוצע המדגם תהיה קטנה יותר**

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

הוכחה:

לפי אי שוויון צבישב מתקיים

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

אם נפעילו על הממוצע $\overline{X}$ נקבל

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

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

נבחר $k \geq \varepsilon\frac{\sqrt{n}}{\sigma}$ ונציבו, נקבל

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

ומכאן נקבל את **חוק המספרים הגדולים:**

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

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

טענה: בדגימת מדגם שגודלו n מתוך ממ X המתפלג נורמלית עם תוחלת μ וסטיית תקן σ יהיה ממוצע המדגם $\overline{X}$ גם הוא ממ המתפלג נורמלית עם תוחלת m וסטיית תקן $\frac{\sigma}{\sqrt{n}}$.

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

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

מתקיים $E[Z_n] = 0$ וכן $Var[Z_n] = 1$.
אזי, יהא $Z \sim N(0,1)$.

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

הערה חשובה: מדגם "גדול מספיק" הוא כזה בגודל שגודלו $n \geq 30$. ורק אם הוא בגודל שגדול מ30 אפשר להשתמש במשפט הגבול המרכזי.

5 הרצאה 5: אמידה סטטיסטית נקודתית

היכן אנחנו כעת נמצאים? לומדים הסקה סטטיסטית. נושא זה מתחלק ל2: אמידת פרמטרים (הרצאה $5-6$) ובדיקת השערות (הרצאה $7-9$). אמידת פרמטרים מתחלקת ל2: בהרצאה זו נדבר על אמידה סטטיסטית נקודתית ובהרצאה הבאה נדבר על אמידת מרווחי בטחון.
היום נדבר על השאלה הבאה: כיצד והאם ניתן להכליל ממצאים במדגם לממצאים באוכלוסייה?

פרמטר: גודל קבוע המאפיין את כל האוכלוסייה.
סטטיסטי: ערך המחושב ע"פ המדגם.
אמידה היא הערכת (שערוך) ערך הפרמטר ע"פ סטטיסטי המדגם.

5.1 אמידה סטטיסטית

ישנן שתי שיטות לעריכת אמידה סטטיסטית.

1. **אמידה נקודתית:** ההכנסה הממוצעת של משפחה בת 4 נפשות היא $11,500$ שקלים בחודש - על סמך המדגם מחושב סטטיסטי אחד.
2. **רווח סמך:** בהסתברות של 80% ההוצאה הממוצעת של משפחה בת 4 נפשות בישראל היא בין 8000 ש"ח ל$16,000$ ש"ח"- על סמך המדגם מחושב טווח של ערכים.

לאמידה סטטיסטית נקודתית ישנן בעיות:

א. בעיה מהותית - הסטטיסטי הוא רק אומדן. כיצד נדע את ערך הפרמטר ביחס לאוכלוסיה כולה?(לא נדע). מדוע מותר להשתמש בהסקה אינדוקטיבית? (לא בקורס הזה).

ב. בעיה מעשית - בעיה כמותית, באיזה סטטיסטי כדי לי להשתמש כדי לאמוד משתנה מסויים? איזה אמדים קיימים ואיזה תכונות יש להם? מה נחשב לאמד טוב?

5.2 בעיית האמידה

נתון: עבור משתנה מקרי $X \sim F$ ונתון מדגם מקרי $x_1,...,x_n$ ב"ת באשר $\forall_{1\leq i\leq n} x_i \sim F$

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

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

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

טרמינולוגיה:

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

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

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

הגדרה: נתון מדגם מקרי $X_1,...,X_n$. אנו רוצים לאמוד את ערכי θ מתוך המדגם. אזי,

1. פרמטר - פונקציה של ערכי האוכלוסייה. יכולה להיות תלויה בפרמטרים לא ידועים.
2. סטטיסטי- פונקציה של ערכי המדגם. אינה תלויה בפרמטרים לא ידועים

דוגמה: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ הוא סטטיסטי. אך $\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$ (השונות) אינה סטטיסטי כי תלויה בμ.

3. אמד $(estimator)$ - סטטיסטי שבעזרתו אומדים פרמטר בלתי ידוע (פונקציה כללית). לדוגמה: הממוצע הוא אמד לתוחלת. **כשאנחנו מחפשים אמד: אנחנו מחפשים נוסחה.**
4. אומדן - המספר עצמו שמציבים בנוסחה (אמד) עבור מקרה ספציפי. התוצאה שקיבלנו עבור האמד במדגם ספציפי (תוצאה ספציפית).

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

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

5. שגיאת האמידה ־ המרחק בין ערך האמד לערך הפרמטר: $\hat{\theta}-\theta$. נשים לב כי את θ איננו יודעים. אז כיצד יעזור לי לחשב ערך אמידה (במדגם ספציפי)? אנחנו נרצה לחסום ככל שניתן את שגיאת האמידה.

6. הטיה של אמד ־ התוחלת של שגיאת האמידה

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

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

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

5.3 תכונות של אמדים

מהו אמד שהוא טוב?

1. **אמד עקבי ־** ככל שהמדגם גדול ההסתברות שהאמד יתכנס לפרמטר האמיתי גדלה. כלומר: $\hat{\theta}\rightarrow_{n\rightarrow\infty}\theta$

2. **חסר הטיה ־** ההטיה של האמד שווה לאפס. כלומר, $E[\hat{\theta}]=\theta \Longrightarrow Bias(\hat{\theta},\theta)=0$. כלומר: אם אמדנו הרבה פעמים, והיו לי שגיאות מהמדד האמיתי בכל אחת מהדגימות אך בתוחלת השגיאות הללו ביטלו אחת את השניה והתקרבנו למדד האמיתי.

תכונות ממוצע המדגם:

עבור תוחלת μ נגדיר את ממוצע המדגם כאמד: $\hat{\mu}=\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i$

האם הממוצע של המדגם הוא אמד טוב לתוחלת?

א. אכן אמד עקבי ־ ככל שהמדגם גדול, ערך האמד $\overline{X}$ מתכנס לערך הפרמטר באוכלוסייה. זה מגיע בדיוק מחוק המספרים השלמים:

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

ב. אכן חסר הטיה ־ ראינו כי $E[\overline{X}]=E[X_i]$ בהרצאה הקודמת (באשר $E[X_i]$ זה תוחלת של מדגם כלשהו), ומכאן שנקבל כי אכן $Bias(\hat{\theta},\theta)=0$.

טענה (**עבור כל התפלגות**)**:** בהינתן מדגם מקרי $x_1,...,x_n$ ב"ת מתוך מ"מ X עם תוחלת μ ושונות σ^2

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

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

הוכחה: של א' :

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

של ב':

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

כעת נדון באמד לשונות עם תוחלת שאינה ידועה. אם אין לנו תוחלת, אולי כדאי להסתכל על הממוצע $\overline{X}$?

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

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

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

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

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

באשר $(*)$ מחייב הסבר: מדובר על $\sum_{i=1}^{n}(X_i-\mu) = (\sum_{i=1}^{n}x_i) - \mu n = n\overline{X} - \mu n = n(\overline{X}-\mu)$

כעת:

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

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

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

ולכן,

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

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

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

טענה: אמד חסר הטיה לשונות באשר התוחלת אינה ידועה הינו:

א. עבור אוכלוסייה (כי יודעים את התוחלת בהכרח):

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

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

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

טענה: אם θ_1 וθ_2 הם אמדים חסרי הטיה עבור θ אזי נעדיף את זה עם השונות הקטנה יותר. את θ_2 המקיים $V(\theta_2) < V(\theta_1)$

יעילות של אמדים: במקרה הכללי - תוחלת ריבועי השגיאות הינה

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

אם θ_1 וθ_2 הם אמדים שאינם חסרי הטיה עבור θ נעדיף את האומד θ_2 המקיים $MSE(\theta_2) < MSE(\theta_1)$

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

5.4 שיטות אמידה

5.4.1 שיטת המומנטים

שיטת המומנטים היא שיטת אמידה על פי פרמטרים המאפיינים התפלגות של אוכלוסיה מסויימת.

נניח משתנה מקרי המתפלג F עבורה ישנם k פרמטרים בלתי ידועים נגדיר **פונקציה מייצרת מומנטים** (mfg). נאמוד את המומנט הk באמצעות ממוצע החזקה הk של התצפיות.

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

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

כל אמד למומנט מחושב כך לפי ערכי $x_1, ..., x_k$ שחושבו במדגם.

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

השיטה אומרת כך:
א. נשווה כל מומנט מסדר k לאומדן שלו במדגם:

$$\mu_1 = g_1(\hat{\theta_1}, ..., \hat{\theta_n})$$

$$\mu_2 = g_2(\hat{\theta_1}, ..., \hat{\theta_n})$$

$$...$$

$$\mu_k = g_k(\hat{\theta_1}, ..., \hat{\theta_n})$$

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

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

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

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

משוואה שניה 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. הרי x_i נתונים לנו וגם n. נקבל

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

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

וכך, מנתונים שידוע שמתפלגים בצורה אחידה, הצלחנו למצוא אמד לa וb הנדרשים.

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

5.4.2 שיטת הניראות המרבית

נניח שהטלתי מטבע מס' פעמים. בכל ההטלות קיבלתי 5 (זה ניסוי ברנולי). לפי מה שזה נראה - נראה כאילו $p(X=5)=1$. נדגיש: **נראה**.

פונקציית הנראות L**:** בהינתן ערך p ניתן לחשב את פונקציית הנראות.
נראה כי בהינתן משתנים שמתפלגים בינומית, נניח ואנחנו יודעים כי ההסתברות שיצא 7 פעמים אותו מספר היא 0.12. כלומר .... $P(k|n,p) =$ נרצה להפוך אותה לפונקציית נראות $L(p|k,n)$.
כיצד נדע איזה ערך ימקסם את L? נגזור אותה לפי p ונשווה לאפס.

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

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

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

נשים לב. מדוע המרנו לln? הרבה יותר קל לגזור כך.
שנית: קיבלנו הוכחה מעניינת לכך ש**השכיחות היחסית היא אמד נראות מרבית עבור פרמטר p בהתפלגות בינומית**.

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

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

ב. נגדיר את לוג פונקציית הנראות

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

ג. נגזור את LL לפי θ ונשווה לאפס למציאת ערך קיצון.
ד. נגזור את LL פעם שנייה לוודא מקסימום.

מינוח. בשאלה של "מצא אנ"מ" עושים את שיטה זו.

מדוע אנ"מ טוב לנו?
א. **עקביות:** ככל שהמדגם גדל, ערך האמד מתקרב לערך הפרמטר.
ב. **אינווריאנטיות פונקציונלית:** אם θ אנ"מ ו g חח"ע אזי גם $g(\theta)$ אנ"מ
ג. **נשים לב - לא ידוע האם האנ"מ הוא חסר הטיה**

לשנן:

מודל תאורטי	התפלגות מ"מ	אמד נראות מירבית – אנ"מ	האם חסר-הטיה – אח"ה
בינומי	X ~ Bin(p)	= X/n	כן
אחידה	X ~ U(1,b)	= max{X1…Xn}	לא
פואסוני	X ~ P(λ)	= X̄	כן
גאומטרי	X ~ G(p)	= 1/X̄	לא
מעריכי	X ~ Exp($\boldsymbol{\theta}$)	עבור $\boldsymbol{\theta}$: = 1/X̄ עבור μ: = 1/X̄	לא כן
נורמלית: - תוחלת - שונות	X ~ N(μ,σ^2)	= X̄ = $\sum_i(X_i - \bar{X})^2 / n$	כן לא

6 הרצאה 6: אמידה סטטיסטית של מרווחי בטחון

היכן אנחנו? לומדים תהליך ניסוי, אנו בחלק של אמידת פרמטרים + בדיקת השערות, נושא לו קראנו סטטיסטיקה היסקית.
ראינו כי סטטיסטיקה היסקית מתחלקת ל2:
א. אמידת פרמטרים: אמידה נקודתית (שיטת המומנטים ושיטת הנראות המרבית) ומרווחי בטחון - **נושא ההרצאה הנוכחית**.
ב. בדיקת השערות - בהמשך.

6.1 רווח סמך של ממוצע המדגם

נתבונן בממוצע המדגם $\overline{X}$ כאמד נקודתי לתוחלת μ. שגיאת האמידה של ממוצע המדגם היא $\overline{X}-\mu$. נשים לב כי שגיאת האמידה לא ידועה לנו, ושגיאת האמידה היא משתנה מקרי בעצמה. תחת תנאים אלו נרצה לבדוק את דיוק האמד. דיוק האמד לא יכול להיות מדד אבסולטי - אלא מדד הסתברותי. לכן נשאל: מהי ההסתברות לך ששגיאת האמידה של האמד תהיה גדולה מ(טווח בטחון)?

נזכר כי לכל $n \geq 30$ עבור משתהה מקרי X בעל תוחלת μ ושונות σ^2 מתקיים $\overline{X} \sim N(\mu, \sigma^2)$ (משפט הגבול המרכזי).

דוגמה.

במדגם שגודלו 25 מתוך מ"מ X המתפלג נורמלית בעל סטיית תקן $\sigma = 10$ ותוחלת μ לא ידועה, מהי ההסתברות שממוצע המדגם יהיה שונה מהתוחלת בלא יותר מ4 יחידות?

נראה כי נתון $X \sim N(\mu, 100)$ וכן $n = 25$, לכן $\overline{X} \sim N(\mu, 4)$ (שונה ב4 יחידות = השונות היא 4)

כלומר, נדרוש

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

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

מה המשמעות של נתון שכזה? אם נזכר בגרף ההתפלגות הנורמלית, המשמעות היא שהשטח שמתחת לפונקציית הצפיפות הוא 95% והזנבות כל אחת 2.5%. כלומר - אם נבצע מדגמים רבים (אנסוף) אז ב95% מהמקרים ממוצע המדגם יפול בטווח זה שנדרש.

נשים לב כי את אי השוויון ממנו התחלנו ניתן להמיר לאי שוויון הבא:

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

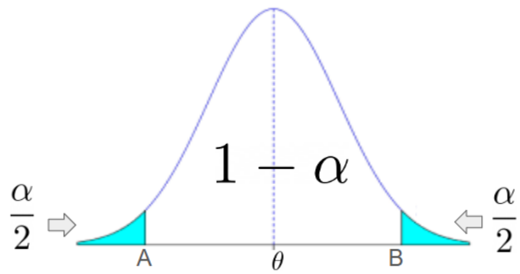
ישנה שקילות בגלל הערך המוחלט. המשמעות לשינוי זה היא - במדגם שגודלו $n = 25$ עבור מ"מ המתפלג נורמלית עם סטיית תקן 4 ותוחלת μ לא ידועה, בהסתברות 0.95 הרווח יכיל את μ. כלומר - מצאנו אינפורמציה חשובה באשר לμ. נקרא לביטוי זה: **רווח הסמך של μ ברמה של 0.95.**

6.2 רווח סמך - הגדרה פורמלית

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

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

כך זה יראה בגרף:



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

6.3 רווח סמך לממוצע עבור התפלגות נורמלית כאשר השונות ידועה כאמד לתוחלת

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

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

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

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

נראה כי בשביל שאגף ימין (ההסתברות) $P(-z < X_Z < z)$ תהיה שווה לרווח הסמך ברמה של $1 - \alpha$:

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

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

דוגמה. אם $\alpha = 0.05$ אזי $\frac{\alpha}{2} = 0.025$ ונקבל $z_{0.025} = 1 - 0.025 = 0.9750$, נלך לחפש ערך זה (0.9750) בטבלת ההתפלגות הנורמלית Z. נראה כי 1.96 מניב הסתברות זו כלומר $Z = 1.96$ ולכן רווח סמך של 95% הוא כאשר $Z = 1.96$

אם נרצה רווח סמך של 90% כלומר $\alpha = 0.1$ ולכן $z_{\frac{\alpha}{2}} = z_{0.05} = 1 - 0.05 = 0.95$ וערך זה מתקבל עבור $Z = 1.645$

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

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

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

דוגמה 2. אורך החיים של נורות מתוצרת מפעל הניצוץ מתפלג נורמלית עם תוחלת μ וסטיית תקן 22. נדגמו רנדומית 16 נורות ונמצא שאורך החיים הממוצע הוא 863. מצא רווח בסמך 90% ל-μ.

פתרון: נראה כי מתקיים $\alpha = 0.1$ ולכן $z_{1-\frac{0.1}{2}} = z_{0.95} = 1.645$ ואם נציב בנוסחת רווח הסמך את הנתונים:

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

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

6.4 רווח 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.

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

כעת, שינינו את התפלגות הדגימה. כיצד יתפלג המ"מ החדש? $Z_{\overline{X}} = \frac{\overline{X}-\mu}{\frac{\hat{S}}{\sqrt{n}}} \sim N(0,1)$ עבור מדגמים $n \geq 30$.

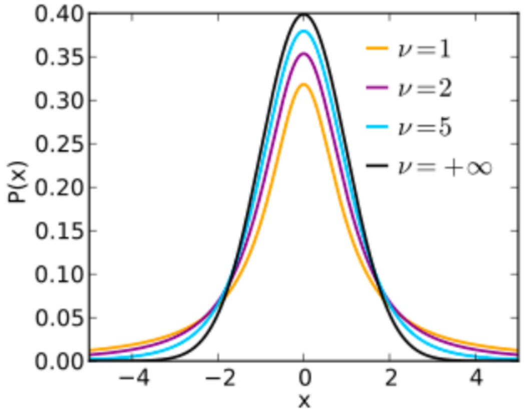
לכן, רווח הסמך ברמת סמך (רמת בטחון) של $1-\alpha$ אחוז עבור μ כאשר השונות אינה ידועה עבור מדגמים $n \geq 30$ היא:

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

ומה באשר למדגמים קטנים? עבור מדגמים קטנים $n < 30$ נסמנו $T_{\overline{X}} = \frac{\overline{X}-\mu}{\frac{\hat{S}}{\sqrt{n}}} \sim t(v)$

התפלגות t:

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



דרגת חופש הינה כמה מספרים יכולים להשתנות באופן חופשי בהינתן הגבלה מסויימת. למשל בהינתן 5 מספרים וממוצע, 4 יכולים להשתנות למה שהם ירצו להיות אך האחרון חייב להתאים את הערך הכולל לממוצע. לכן דרגת החופש של 5 המספרים הוא 4.

התפלגות זו מתקרבת לZ (נורמלית) ככל שיש יותר דרגות חופש.

כל אמד מוריד לנו דרגת חופש אחת, כיוון שתלויים עוד ועוד. אנו נסמכים על חישוב הממוצע, ולכן דרגת החופש v היא גודל המדגם פחות אחד. כלומר,

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

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

דוגמה. במדגם בגודל $n = 9$ מאוכלוסיה מתפלגת נורמלית נמצא הממוצע 114. האומדן לסטיית התקן באוכלוסיה הינו 12. מצא רווח סמך לתוחלת ברמת סמך של 95%.

פתרון: נתון לנו $n = 9 < 30$ לכן נשתמש בטבלת התפלגות t. וכן $v = 9 - 1 = 8$ ו$\overline{X} = 114$. כמו כן $\alpha = 0.05$ וכן $t_{\frac{\alpha}{2}}(8) = t_{0.025}(8) = 2.306$ באשר ערך זה מהטבלה. מכאן נקבל

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

סיכום:

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

- **ממוצע המדגם** הוא אמד עקבי וחסר הטייה ל**תוחלת**
- בחישוב **רווח סמך** עבור התוחלת ברמת סמך α ו**בשונות 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 **רווח סמך** עבור התוחלת ברמת סמך α בשונות **לא** ידועה מדגמים **גדולים**:

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

- בחישוב **רווח 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, מרווח הטעות הינו $ME = z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ ומתקיים $\overline{X} - ME < \mu < \overline{X} + ME$.

7 הרצאה $8+7$: הסקת סטטיסטית - בדיקת השערות

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

7.1 השערת האפס וההשערה האלטרנטיבית

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

נסמן את ההשערה הראשונית - **השערת האפס ב**H_0, ואת ה**השערה האלטרנטיבית ב**H_1.
לרוב H_0 היא שאין הבדל בין המדגמים וH_1 היא שיש הבדל מסויים (בממוצע, בהתפלגות, וכו').

לדוגמה: אם נרצה לדון בשאלה "האם גובהם הממוצע של גברים אסיאתים גדול משל שאר הגברים באוכלוסיה"? אזי אם נסמן μ_{Asian} כממוצע הגובה של גברים אסיאתים ו$\mu_{Non-Asian}$ כממוצע הגובה של גברים שאינם אסיאתים אזי:

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

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

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

איך נחליט איזו מההיפותזות נכונה?
קיימות שתי גישות בסיסיות לקבלת החלטה:
א. חישוב הסבירות של השערת האפס H_0.
ב. גישת קבלת החלטות (מזעור השגיאה).

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

7.2 סוגי השגיאות

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

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

שגיאה מסוג 1: דחיית H_0 בטעות - יש לה שמות נוספים כגון $False\ positive$, α, ל-$1-\alpha$ קוראים $true\ negative$.

שגיאה מסוג 2: קבלת H_0 בטעות - יש לה שמות נוספים כגון $False\ negetive$, β, ל-$1-\beta$ קוראים $true\ positive$.

החלטת החוקר (על סמך המדגם)	H_0 **נכונה** במציאות	H_0 **שגויה** במציאות
אי-דחיית H_0	החלטה נכונה (**True Negative**)	**שגיאה מסוג II**
דחיית H_0	**שגיאה מסוג I**	החלטה נכונה (**True Positive**)

7.3 חישוב הסבירות של השערת האפס

נניח שאנו רוצים לבדוק אם ממוצע המדגם של גבהים $(\overline{X})$ שונה מהממוצע של האוכלוסיה (μ_0) בגבהים. נרצה שמדד עבור הסבירות של השערת האפס יהיה קטן יותר ככל שההבדל בין μ_0 ל-$\overline{x}$ גדול יותר (שהסבירות שהשערת האפס נכונה - יהיה לא גבוה אם הם רחוקים). אפשר לחשב מדד כזה ע"י פונקצייה של סטיית התקן סביב הממוצע:

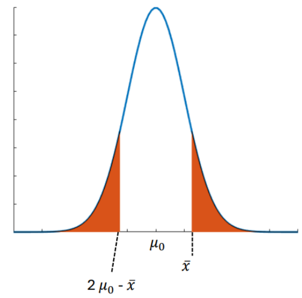
$$\frac{\overline{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$$

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

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

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

משמעותו של ה$P-value$**:** השטח שצבוע הוא ה$P-value$. תחת H_0 (כלומר, אם ההשערה נכונה) ההסתברות ש$\overline{x}$ גדולה או שווה מזו שנצפתה היא $p-Value$. כלומר, ככל שה$p-Value$ יותר קטן, פחות סביר שH_0 נכונה.



צדדים ב$P-value$**:**

א. מבחן $right-tailed$: אם אנחנו חושדים מראש (לפני שראינו את הנתונים) כי $\bar{x} < \mu_0$.

ב. מבחן $left-tailed$: אם אנחנו חושדים מראש (לפני שראינו את הנתונים) כי $\bar{x} > \mu_0$.

ג. מבחן דו צדדי $(two-tailed)$: ברירת המחדל. חושדים שהם שווין. נשים לב שהנוסחאות לחישובו ישתנו במבחן דו צדדי.

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

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 שאם ה$P-value$ יהיה קטן ממנו נדחה את השערת האפס (למשל, סיפים מקובלים הם $0.001, 0.05, 0.01$)

ב. לחשב את ה$P-value$ של המדגם.

3. לדחות את השערת האפס אם ה$P-value$ קטן מהסף שקבענו. לאחר שלב זה אין משמעות לגודלו של ה$P-value$.

נניח שה$P-value$ גדול מהסף שהוחלט מראש. האם זה גורר שH_0 נכונה? לא! זה רק אומר שאי אפשר לדחות את השערת האפס. מדוע? אין מספיק נתונים, או שהמידע רועש מדי.

צורה נוספת לחשוב על ה$P-value$**:** ההסתברות לקבל את המדגם (הנתונים) שקיבלנו, בהנחה שהשערת האפס היא נכונה.

הגדרה: אם החלטנו לדחות את השערת האפס כיוון שH_0 היה נמוך מהסף שקבענו, נאמר שהשערת האפס נדחתה באופן מובהק סטטיסטית.

7.5 תיקון למבחנים מרובים

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

מספק המבחנים. לשם כך נשתמש בתיקון $Bonferroni$ -
יש לקבוע את הסף ל$P-value$ כ$\frac{\alpha}{n}$ באשר α הוא הסף למבחן בודד, וn הוא מס' המבחנים. תיקון זה הוא שמרני, ישנם אלטרנטיביות.
כלומר, אם הסף הקודם היה α נאמר כי הסף החדש הינו:

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

ואז, נדחה עבור מבחן ספציפי אמ"מ $P-value < \alpha'$

7.6 חישוב גודל המדגם הדרוש

קודם לכן ראינו כי:

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

ולכן -

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

וכך אפשר (תחת הנחות על ההתפלגות של X) לחשב כמה דגימות דרושות כדי לגלות הבדל משמעותי סטטיסטית ברמת מובהקות נתונה.

7.7 מספר מדגמים

עד כה - התייחסנו למצב שבו יש מדגם אחד שמשווה לערך תאורטי בודד. מה קורה באשר ישנם שתי מדגמים?

מדוע זה שונה? במצב זה, צריך לקחת בחשבון את הפרמטרים של שתי ההתפלגויות ואת השכיחות היחסית של שתיהן.

עד כה - דנו בשאלה כיצד אפשר לשלוט בשגיאה מסוג 1: ההסתברות לדחות את השערת האפס בטעות.

עוצמת המבחן: השם שניתן ל$1-FN$, כלומר: המשלים של הסיכוי לשגיאה מסוג 2- ההסתברות לקבל את השערת האפס בטעות (β). כלומר, עוצמת המבחן היא ההסתברות לדחות את השערת האפס כשהיא באמת שגויה.

עוצמת המבחן, היא ההסתברות לדחות 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 בדיוק השגיאה מסוג 2.

כיצד מחשבים את עוצמת המבחן? ראשית מחשבים את $\beta = \phi(Z) = \phi(\frac{C-\mu_1}{\sigma\setminus\sqrt{n}})$ בעוד C הוא הערך הקריטי (כתלוי במבחן הערך שקובע מתי נדחה את H_0 לפי רמת מובהקות α), μ_1 הוא הממוצע האמיתי שלכאורה נכון תחת H_1. ולאחר מכן, מחשבים את $1-\beta$.

עוצמת המבחן תלויה: במבחן בו משתמשים, ברמת המובהקות, בנתונים (גודל המדגם וכו'), גודל האפקט שאותו מנסים לזהות (נדבר על מושג זה בהמשך).

עוצמת מבחן גבוהה תתקבל (באופן כללי) אם יש לנו - שונות נמוכה בנתונים, מדגם גדול, גודל אפקט גדול, דרישות נמוכות מרמת המובהקות.

במילים אחרות: עוצמת המבחן היא ההסתברות לתפוס את ההבדל באשר הוא אכן קיים. כלומר - לדחות נכונה את H_0 אם H_1 נכון. אם יש לנו רופא שטוען H_0 התרופה עובדת וH_1 התרופה אינה עובדת. אזי, עוצמת המבחן היא ההסתברות לומר כי התרופה לא עובדת כאשר התרופה באמת לא עובדת. לכן, זה המשלים לשגיאה מסוג 2: הסתברות שנטען כי H_0 נכונה בעוד H_1 היא הנכונה. לכן, אם עוצמת המבחן גבוהה, סיכוי טוב שנמצא את מה שאנחנו מחפשים אם הוא באמת שם.

7.8 גודל האפקט (הd של כהן)

מוגדר כ

$$d = \frac{\overline{x_1} - \overline{x_2}}{s}$$

באשר s היא סטיית התקן. קיימים מדדים נוספים לגודל האפקט. הוא מתאר את העוצמה או הגודל של הקשר או ההבדל שזיהינו במדגם, באופן שאינו תלוי בגודל המדגם.
מובהקות סטטיסטית ($P-value$)**:** אומרת לנו האם יש הבדל/אפקט (האם הוא אמיתי או מקרי).
גודל האפקט (d)**:** אומר לנו כמה גדול ההבדל/האפקט.

7.9 סיכום חשוב

במציאות ישנם 2 מצבים בלבד. או שH_0 נכונה, לא קיים אפקט בכלל. או שH_1 נכונה, קיים אפקט כלשהו ומשהו השתנה.

אבל איננו יודעים מהו המצב האמיתי. לפיכך,
המבחן הסטטיסטי הוא למעשה קו החלטה.
אם התוצאה קיצונית מספיק $\Longrightarrow$ יש אפקט $\Longrightarrow$ דוחים את H_0
אם התוצאה לא קיצונית $\Longrightarrow$ אין אפקט $\Longrightarrow$ לא דוחים את H_0
אבל הקו הזה לא מושלם, קו ההחלטה יכול לטעות. נתבונן בשני עולמות מקבילים:

עולם ראשון - H_0 נכונה (אין אפקט): בעולם זה, אם נריץ את הניסוי הרבה פעמים ברוב הפעמים נקבל תוצאות רגילות ונגיד שאין אפקט. בחלק קטן מן הפעמים, ב5% מהם (אם לקחנו $\alpha = 0.05$) אנחנו נקבל תוצאה קיצונית ונגיד בטעות "יש אפקט" בעוד עודנו יודעים שאין אפקט שכזה: זו בדיוק שגיאה מסוג ראשון.

עולם שני - H_1 נכונה (יש אפקט): בעולם זה, אם נריץ את הניסוי המון פעמים: ב$power\%$ מהפעמים $(1-\beta)$ אנחנו נקבל תוצאה קיצונית ונגיד: יש הבדל, בעוד אכן יש הבדל. זו בדיוק עוצמת המבחן.

ב$\beta\%$ מהפעמים, נקבל תוצאה לא מספיק קיצונית ונגיד בטעות: אין אפקט. זו בדיוק שגיאה מסוג שני.

הערה חשובה: אם α קטן יותר, המשמעות היא שאזור הדחייה הפך להיות קטן יותר - כיוון שאנחנו דוחים את H_0 רק בערכים מאוד קיצוניים. ולכן, במקרה זה אזור הקבלה הופך לגדול יותר.

7.10 מבחני השערות במדגמים גדולים

מהו מדגם גדול? מקובל לומר שמדגם 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 מובהקות סטטיסטית. אם המדגם קטן, יש לבצע תיקון למדדים.

מקרה מפורסם שכזה: בהינתן n דגימות בלתי תלויות של משתנה נורמלי, נחשב:

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

הנתונים מתפלגים בהתפלגות t עם מס' דרגות חופש השווה ל$n-1$. 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 וΓ היא פונקציית גאמא. לשמחתנו, באשר n גדול התפלגות t שואפת להתפלגות נורמלית.

7.11 מבחני השערות

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

7.11.1 דוגמה ראשונה: ממוצעים

במצב זה, יש לנו טענה על האוכלוסיה: הממוצע באוכלוסיה הינו $\overline{x}$. איננו יכולים לבדוק את כל האוכלוסיה. אז מה נעשה? נדגום מדגם קטן ונבדוק מה הממוצע שם. השאלה שנרצה לשאול, האם הממוצע שמצאנו באוכלוסיה סותר את הטענה המקורית או שלא? או במילים אחרות ־ האם ההבדל שמצאתי בין מה שטענו נבע בגלל רק מקרה בודד או רעש סטטיסטי או ממש משמעותי ואז הטענה המקורית אודות הממוצע הייתה שגויה.

נסמן:
א. $\overline{x}$ ממוצע המדגם
ב. μ תוחלת האוכלוסיה
ג. σ סטיית התקן של האוכלוסיה
ד. n גודל המדגם

המשתנה המתוקנן יהיה $Z = \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ והוא אומר למעשה ־ כמה יחידות רעש אני רחוק מהממוצע המוצהר? וכעת, אם נרצה רמת מובהקות של α בטענה,
אזי לא נדחה את השערת האפס אם:

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

וכן נדחה אותה אם הדבר אינו מתקיים.

דוגמה לממוצעים. טענה: H_0: הגובה הממוצע של גברים בישראל $\mu_0 = 175$. אנחנו נדגום $n = 100$ גברים אקראיים. נקבל כי $\overline{x} = 177$ של האנשים שדגמנו. נבחין כי $175 \neq 177$. מדוע זה קרה? אפשרות ראשונה: זה קרה במקרה, אם הייתי דוגם עוד מאה הייתי מקבל $\mu = 173$. במקרה זה, הטענה המקורית עדיין נכונה ולא דוחים את H_0. אפשרות ב': ההבדל גדול מדי בשביל להיות מקרי. כיצד מחליטים? המבחן הסטטיסטי מחליט. הוא שואל: מה הסיכוי לקבל ממוצע של 177 במדגם באשר הממוצע באוכלוסיה הינו 175? הוא קובע רמת מובהקות, ומחשב בהתאם.

7.11.2 דוגמה שנייה: הצלחות במדגם

כעת אנו במצב בו שיעור ההצלחות/ כן/ מסכימים באוכלוסיה הינו $p = \mu$. כלומר, p הוא יחס ההצלחות במדגם (למשל, מס' הפעמים שניסוי הצליח מתוך כלל הניסויים). דגמנו מדגם, וקיבלנו שיעור $P = \mu_p$. כעת, נרצה לבדוק האם ההבדל בין P לp הוא משמעותי?

כיוון שמדובר במשתנה בינומי, $\sigma_p = \sqrt{\frac{p(1-p)}{n}}$. נרצה לתקנן, המשתנה המתוקנן הינו:

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

וכעת, 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 שבבחירות הקרובות, $p = 0.4$ הוא אחוז התמיכה במועמד א' בבחירות. בסקר, שאלנו $n = 500$ אנשים ומתוכם $P = \frac{230}{500} = 0.46$ תומכים במועמד. נרצה לבדוק: האם 0.4 שונה משמעותית מ0.46? נחשב את Z לפי הנוסחה ונקבל $Z = 2.74$. כעת, $2.74 > 1.96$ ברמת מובהקות של 5% ולכן דוחים את H_0 ברמת מובהקות 5%.

חשוב: אם מתקיים $np > 10 \wedge n(1-p) > 10$ אזי אפשר לומר כי פורפורציית המדגם מתפלגת נורמלית כלומר $\hat{p} \sim N(p, \frac{p(1-p)}{n})$ דהיינו $\mu = p, \sigma^2 = \frac{p(1-p)}{n}$. כעת נוכל להשתמש ב$Z$ בשביל לדעת את פונקציית הצפיפות המצטברת של משתנה בינומי.

7.11.3 דוגמה שלישית: הפרשים בין ממוצעים

עד כה השווינו מדגם אחד לטענה תאורטית. 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. כלומר: בדקנו שתי קבוצות, אסיאתים ולא אסיאתים. הממוצע של האסיאתים בגובהם הינו 174 ס"מ והממוצע של הלא אסיאתים בגובהם הינו 177 ס"מ. אנו שואלים - האם ההפרש של 3 בניהם היה מקרי, או שפשוט הטענה אודות ממוצע הגבהים שלהם היא שהם לא שווים. נפרמל,

יהיו שני ממוצעים μ_1, μ_2. בהינתן כי H_0 היא $\mu_1 = \mu_2$ וH_1 היא ש$\mu_1 \neq \mu_2$ (הממוצעים שונים). לכן, נמיר את המבחן למבחן של הבדל הממוצע מאפס, אותו אנחנו יודעים 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 בהתאמה $\overline{x_1}, \overline{x_2}$ ואת גדלי המדגמים n_1, n_2 בהתאמה. סטיית התקן של הפרשי המדגמים:

$$E^2_{S1-S2} = E((X_1 - X_2)^2) =^* E(X_1)^2 - 2E(X_1X_2) + E(X_2)^2 = E(X_1)^2 + E(X_2)^2$$

$*$ שכן הגורם $-2X_1X_2$ יוצר תוחלת 0 כיוון שהגורמים בלתי תלויים. לכן, שגיאת התקן של ההפרש בין המדגמים הינה:

$$\sigma_{S_1-S_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

המשתנה מתוקנן של ההפרש בין הממוצעים הינו:

$$Z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

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

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

הערה חשובה. מאוד. אם $\mu_1 = \mu_2 + 2$ כהשערת האפס, אזי כמובן ש$\mu_1 - \mu_2 = 2$ ולא אפס ולכן הנוסחה בהתאם תהיה $Z = \frac{\overline{x_1} - \overline{x_2} + 2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$.

כעת, אם Z יהיה מאוד גדול אזי ההפרש ביניהם כנראה לא היה אפס ונדחה את השערת האפס ולכן נאמץ את ההשערה האלטרנטיבית H_1. במקרה של $\alpha = 0.05$ נקבל כי $Z_\alpha = 1.96$. נניח וZ שלנו יצא $Z = 60$, אזי $1.96 <<<<< 60$! ההפרש עצום ולכן בהכרח נדחה את השערת האפס.

באופן דומה, אם נסתכל על הפרש בין הצלחות בין קבוצות, כלומר האם שיעור ההצלחות בקבוצה 1 שונה משמעותית משיעור ההצלחות בקבוצה 2?

נבחין כי אם בוצעו n_1, n_2 ניסויים שהצליחו בהתאמה בx_1, x_2 פעמים אזי $p = \frac{x_1+x_2}{n_1+n_2}$, ומכאן:

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

ונקבל כי

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

דוגמה. נניח כי $n_1 = 500$ גברים ראו פרסומת ו$x_1 = 150$ לחצו עליה. אזי $P_1 = \frac{150}{500} = 0.3$. וכן $n_2 = 600$ נשים ראו פרסומת ו$x_2 = 120$ לחצו עליה אזי $P_2 = \frac{120}{600} = 0.2$. כיצד נבדוק את ההשערות? השערת האפס הינה $P_1 = P_2$ והאלטרנטיבית היא שהם שונים. לפיכך, נחשב את $p = \frac{x_1+x_2}{n_1+n_2} = 0.245$. מכאן נחשב את סטיית התקן שתצא $\sigma_{p1-p2} = 0.026$. נחשב את Z לפי הנוסחה מלמעלה ונקבל כי $Z = 3.85$. אם $\alpha = 0.05$ אזי כיוון ש$1.96 < 3.85$ 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 $(x_1, y_1), ..., (x_n, y_n)$ - 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: $D_1, ..., D_n$ של ההפרשים בין התוצאות. נבחין כי כעת הם בלתי תלויות. (**תמיד להזכר בדוגמאות עם ההצבעות לטראמפ. (x_1, y_1) זה ההצבעות לטראמפ בפלורידה ב$2016, 2020$ בהתאמה. ולכן אם נסתכל על ההפרש בין ההתאמות, נקבל סדרה חדשה של 50 המדינות 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 המתוקנן יהיה:

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

שכן כעת התוחלת היא 0 תחת השערת האפס (אותה אנו מניחים במבחני ההשערות). נשים לב כי הנחת היסוד של המודל היא שההפרשים אכן מתפלגים נורמלים.
הערה. נבחין כי בכוונה סימנו t, Z זה כאשר אנו יודעים את σ סטיית התקן האמיתית וt זה באשר אנחנו מעריכים את σ מתוך המדגם S. אם $n > 30$ נוכל להשתמש בt, ואם $n \leq 30$ נהיה חייבים להשתמש בטבלת t.

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

7.11.5 מבחנים אפרמטריים: מבחן χ^2

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

עד כה עבדנו עם מבחנים פרמטריים, הנחנו הנחות על הנתונים ־ הנתונים מתפלגים נורמלית, התוחלת היא μ, סטיית התקן היא σ וכדומה. מבחני Z וt הם מבחנים פרמטריים. אם כן, אלו מבחנים שרגישים הרבה יותר לנתונים חריגים, דורשים שהנתונים יהיו נורמליים ולא עובדים טוב על נתונים קטגוריאליים.

מבחנים אפרמטריים הם מבחנים שאינם מניחים הנחות על התפלגויות הנתונים. לכן: פחות רגישים לנתונים יוצאי דופן, אין צורך לבדוק את התפלגות הנתונים והם עובדים הן על נתונים אורדינליים והן על נתונים קטגוריאליים. מה החסרונות? הם חזקים פחות ממבחנים פרמטריים ־ דורשים מדגם גדול יותר.

מבחן χ^2 הוא מבחן אפרמטרי שעונה על השאלה הבאה: האם ההתפלגות שאני רואה בנתונים תואמת להתפלגות שציפיתי לה?

למשל: זריקת קוביה ־ יש לך קוביה ואתה זורק אותה 600 פעמים. אם הקוביה הוגנת, אתה מצפה לראות בממוצע כל מספר 100 פעמים. אם בפועל קיבלת רק 107 עבור 1 ו93 עבור 2 למשל, האם הבדל זה קרה בטעות במקרה רעש או בכוונה והקוביה אינה הוגנת?

הרעיון כזה, נמדוד כמה רחוקה ההתפלגות שראינו מההתפלגות שציפינו. לשם כך, נתבונן 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 שצפינו אליו ־ $Expected$. במילים אחרות הנוסחה אומרת: חשב את הפער, חלקי מה שציפינו אליו. מדוע מעלים בריבוע? בשביל שכיוון הפער לא יהיה חשוב. מדוע מחלקים בE_i? בשביל ליצור הפרש יחסי. אם ההפרש יצא 20 אך קיבלנו 20 וציפינו ל40, זה לא אותו דבר כמו שציפינו ל120 וקיבלנו 100.

לוקחים דוגמה של קוביה למשל, עם נתונים ידועים. מחשבים ומקבלים כי $\chi^2 = 2.58$. מס' דרגות החופש הינו $k-1$, כיוון שישנם 6 ערכים (אפשריים של מספרים בקובייה) אזי מס' דרגות החופש הינו 5. שכן, אם קבענו 5 ערכים עבור כמה הטלות יצא בהם $1, 2, 3, 4, 5$. אזי ברור שכיוון שסכום ההטלות הוא 600 ידועה התוצאה האחרונה. כעת, מחפשים בטבלת χ^2 את $\alpha = 0.05$ (רמת וודאות) וכן מס' דרגות חופש $v = 5$. מקבלים כי $C = 11.07$ (ערך קריטי). כיוון שאצלנו $\chi^2 = 2.58 < 11.07$ אנחנו לא דוחים את H_0 והקוביה הינה הוגנת.

באופן כללי עבור מבחן χ^2: אם נקבל $C > \chi^2$ אזי לא נדחה את H_0 ואם נקבל $C < \chi^2$ כן נדחה את H_0 במבחן ההשערות.

וכן, אם χ^2 קטן אזי זה קרוב למה שצפינו, יכול להיות רעש רגיל ולכן לא דוחים. אם χ^2 גדול אזי זה רחוק ממה שצפינו ־ לא יתכן שזה רק רעש, ולכן דוחים.

הערה חשובה. נבחין כי χ^2 בהכרח חיובי, לכן אם $\chi^2 = 0$ זו התאמה מושלמת, בדיוק מה שרצינו. בכוונה מדובר בχ^2 ־ שהערך יהיה תמיד חיובי. הערך הקריטי שהרי הוא קו הגבול הוא יקבע האם נדחה או לא נדחה.

7.11.6 מדידת הפרשים א־פרמטרית במדגמים בלתי תלויים: מבחן $Mann$ $Whitney\ U\ TEST$

מדוע אנחנו זקוקים למבחן נוסף? עד כה, בשביל להשוות שתי קבוצות השווינו ממוצעים עם מבחן t.

אבל מבחן t דורש התפלגות נורמלית, מדגמים גדולים ונתונים רציפים. מה אם הנתונים לא נורמליים? המדגם קטן? יש ערכים קיצוניים? ישנם נתונים אורידנליים (דירוג, לא מספרים)? לשם כך נשתמש במבחן הא־פרמטרי הבא שלא דורש הנחות על ההתפלגות או הפרמטרים.

המדגם בודק האם החציונים של שני התפלגויות זהים. משווים את החציונים של שתי ההתפלגויות. נתונות דגימות x, y מהתפלגויות אורדינליות רציפות F_x, F_y בהתאמה. כיוון שההתפלגויות רציפות $P(x=y)=0$. **ההשערות:**

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

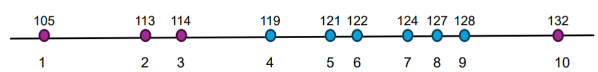
$$H_1 : P(x>y) \neq P(y>x)$$

כלומר, אנחנו בודקים האם החציונים של ההתפלגויות שווים. אם הם שווים אזי ההסתברות למישהו מעליו או מתחתיו זהה.

נתבונן בדוגמה: האם הטיסות מישראל לאתונה לוקחות יותר זמן מהטיסות מאתונה לישראל?

זמני טיסה, בדקות

TLV-ATH	ATH-TLV
121	113
122	114
119	132
128	105
124	
127	



המבחן מציע את הרעיון הבא: נסדר את הנתונים על קו ישר" באשר נזכור לכל נקודה מהיכן היא הגיעה ־ מאיזו דגימה וכן נסדרם בסדר עולה. נבחין כי אם H_0 נכונה (אין הבדל) אזי הם צריכים להיות מעורבבים (כי החציונים שלהם שווים), ואם H_1 נכונה אזי 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 שאם אין הבדל, אזי הדירוגים צריכים להיות מעורבבים באקראי ואם יש הבדל הדירוגים יהיו מקובצים.
נבחר את סך המיקום הקטן יותר ונשאל, היכן הוא ביחס לכל הסדרים האפשריים? נסתכל על $R_1 = 16$, זה סכום די קטן ויש לקבוצה 1 ערכים בעיקר קטנים.
ישנם $\binom{10}{4}$ סדרים אפשריים (בוחרים 4 לסגולים, וכל השאר כבר יסתדרו בהתאם) לחלק את הדירוגים. וכעת נסתכל על סכום הסדרים האפשריים. נראה היכן עומד R_1 ביחס לסדרים. במקרה שלנו, אם נחשב (המחשב יחשב) נקבל כי רק 13% מהפרמוטציות יותר קטנות (כלומר, רק 13% מהפרמוטציות מניבות $R_1' < R_1 = 16$). ולכן זה לא מובהק סטטיסטית ־ זה לא דוחה את השערת האפס. אם היינו מקבלים שמס' הפרמוטציות שקטנות יותר הוא קטן מ־5% אזי זה היה מובהק סטטיסטית.

ובאופן כללי -

מדוע אנחנו צריכים משתנה U? R הוא תלוי בגודל המדגם. ולכן, אנחנו מנרמלים את R בהינתן לגודל המדגם. שוב, R_y זהו סכום הדירוגים של הקבוצה הy וn_y זהו מס' התצפיות בקבוצה הזו. אנחנו ננרמל אותו על ידי כך שנוריד ממנו את הפרמוטציה המינימלית האפשרית: זו שנקח בה את המיקומים $1, 2, 3, ..., n$ (הסכום יהיה הכי 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 מעל המינימום. אם נחזור לדוגמה הקודמת, מתקיים $U_1 = 16 - \frac{4 \times 5}{2} = 6$, כלומר אנחנו 6 נקודות מעל המינימום. נבחין כי U המינימלי הוא זה שהסדר בו הוא המינימום ולכן $U = 0$. אם U בינוני - הקבוצה מעורבבת ואם U גדול - הקבוצה מקובצת בחלק העליון.

ככל שהמדד הזה יותר גדול, ההבדל פחות מובהק.

ישנה טבלת U, לא נשתמש בה. תמיד יכתבו לנו שמדגם מספיק גדול ולכן נשתמש בקירוב הבא: נבחין כי במדגם קטן יכולנו לחשב את מספר האפשרויות לסידור, אך במדגמים עם n גדול זה מספר אסטרונומי. לפיכך, כשהמדגמים מספיק גדולים $n_1, n_2 > 10 - 20$ נשתמש במבחן Z רגיל באשר החישוב יהיה כך:

הממוצע יסומן m_U ויחושב כך: $m_U = \frac{n_1 n_2}{2}$

סטיית התקן תהיה $\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$

וכן המשתנה הגאוסיני יהיה $Z = \frac{U - m_U}{\sigma_U}$

דוגמה. האם גברים גבוהים יותר מנשים? 50 גברים ו60 נשים. אחרי סידור כל 50 ו60 האנשים יחד קיבלנו כי $R_1 = 3200$ (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 הקריטי באשר $\alpha = 0.05$ הינו $Z = 1.96$, במבחן דו צדדי נדחה את H_0 כי $2.55 > 1.96$. כלומר: גברים גבוהים יותר מנשים. **נבחין כי אינטואטיבית אכן הגברים גבוהים יותר שכן אם זה לא היה המצב המיקום הממוצע לגבר היה $\frac{\sum_{i=1}^{110} i}{110} = 55.5$ אך זה גדול מהממוצע - כלומר הגברים מקבלים**

דירוג גבוה יותר מהממוצע הכללי.

הערה חשובה. אם יש שני ערכים זהים הם יקבלו את אותו ערך הדירוג.
הערה שנייה. עלינו לחשב ערכי U לכל הקבוצות, ולבחור את המינימלי ביניהם ואיתו ללכת לטבלה.

שלבי המבחן:

א. שלב ראשון: סדר אותם לפי סדר עולה.
ב. חשב את הראנקים R_i
ג. הגדר את U_i המתאימים
ד. הגדר $U = min\{U_1, U_2\}$
ה. 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 מזווג (מבחן $Wilcoxon\ Singed{-}Rank\ Test$)

מדוע אנחנו צריכים מבחן נוסף? לקבוצות בלתי תלויות ישנו מבחן פרמטר t, ומבחן אי פרמטרי של $Mann - Whitney - U$. לקבוצות תלויות: יש לנו את מבחן t לזוגות וכעת נלמד מבחן א־פרמטרי לזוגות.

נשתמש במבחן זה כאשר אותם אנשים נמדדו פעמיים (לפני, אחרי), זוגות מותאמים (תאומים, אותו אדם), הנתונים לא נורמליים או יש חריגות, מדגם קטן.

כלומר, נניח שיש שני מדגמים מזווגים (x_i, y_i). נרצה לבדוק אם החציון של x_i שונה מהחציון 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 אותם אנשים בדיוק, שרצים כל אחד פעמיים: פעם אחת בכל שנה. אי אפשר לטעון שאלו זוגות בלתי תלויים ־ יש תלות בין כל שניים. 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 ב2024 ושלילי $-$ אומר כי רץ מהיר יותר ב2024.

מסדרים את הערכים לפי גודל ההפרש ־ ללא סימן. לאחר מכן, לכל ערך מחזירים את הסימן שלו ומקבלים סידור שלהם באופן מדורג.

מחשבים את w^+ ו־w^-: שתי קבוצות המייצגות בהתאמה את האינדקסים המדורגים של האיברים שסימנם חיובי, ואיברים שסימנם שלילי.

מתבוננים ב$W = min\{w^+, w^-\}$ ־ נרצה לדעת את המיעוט המשמעותי.

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

הולכים אל הטבלה: $Wilcoxon\ Singed-Rank$, מתבוננים במס' הזוגות $n=10$ וכן $\alpha=0.05$ ומקבלים ערך קריטי $C=8$. 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 כך 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 $D_i=X_i-Y_i$ לכל זוג (X_i,Y_i)

ב. הסר אפסים: אם $D_i=0$ עדכן את n בהתאם והוצא את הזוג מהמדגם.

ג. לכל זוג קח את ההפרש $|D_i|$ בערך מוחלט.

ד. סדר את $|D_i|$ בסדר מהקטן לגדול.

ה. תן דירוג $Rank: 1,...,n$ 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 בהתאמה. הגדר $W=min\{w^{+},w^{-}\}$

ח. אם $n\leq 20$: השווה את W לערך הקריטי (n,α) בטבלת $Wilcoxon$:

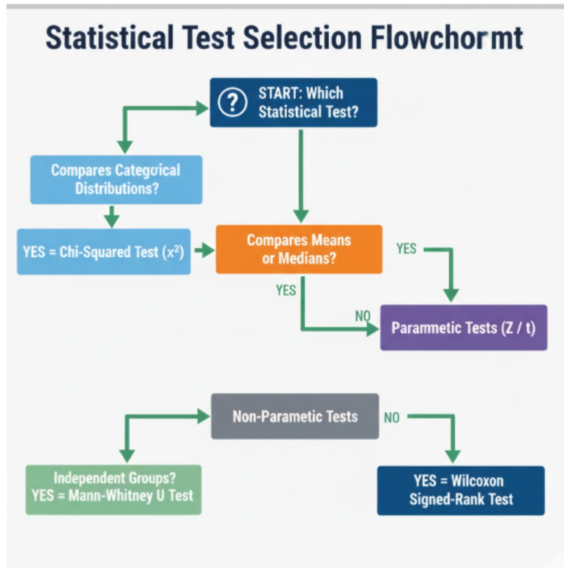
1. אם $W<C$ דוחים את H_0
2. אם $W\geq C$ לא דוחים את H_0

ט. אם $n>20$ (מדגם גדול) חשב את $Z=\frac{W-\mu_W}{\sigma_W}$

1. השווה לערך הZ עם רמת ודאות α. אם בתחום - אל תדחה, אחרת תדחה את השערת האפס.

7.12 סיכום מבחני השערות

באיזה מבחן להשתמש?



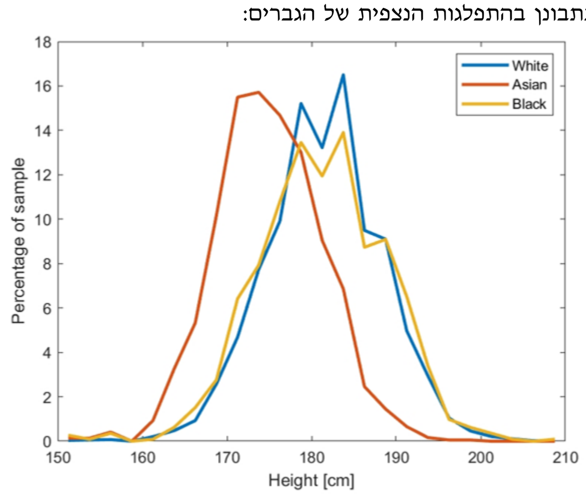
8 הרצאה 9: $Anova\ (Analysis\ of\ Variance)$

עד כה, דיברנו על בדיקת השערות במשתנה אחד (או שניים). כעת נדבר על בדיקת השערות באשר יש לנו יותר מ—2 קבוצות. בתחילה, נדבר כיצד נעשה זאת במשתנה אחד ובהמשך נרחיב ליותר ממשתנה אחד. (כלומר, נדבר על בדיקת השערות של משתנה אחד ביותר מ2 קבוצות, ובהמשך נרחיב לבדיקת השערות של כמה משתנים ביותר מ2 קבוצות).

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

לדוגמה: האם גברים ממוצא אסיאתי שונים בגובהם מגברים אחרים? נבחין, כי בנתונים ישנם הרבה קבוצות נוספות: אנשים לבנים, אסיאתים, שחורים, הודים, אינדיאנים וכן הלאה. כעת ־ ניתן לחלק את הנתונים ליותר מ2 קבוצות ולשאול שאלה אחרת: האם גברים שונים בגובהם כתלות בקבוצה האתנית שאליה הם משתייכים?

נתבונן בהתפלגות הנצפית של הגברים:



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

8.1 הנחות יסוד למבחן אנובה (בסיסי)

מטרה של המבחן: להשוות בין יותר מ2 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: גזע שהם מגיעים מהם, לא משתנה רציף כלשהו!

ה. המשתנה שנמדד הוא רציף.

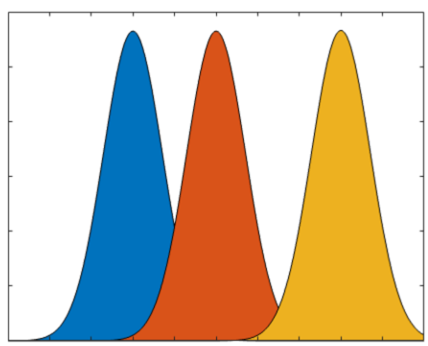
בהמשך, ננסה לוותר על חלק מההנחות. אך: עם ההנחות חיינו יהיו פשוטים הרבה יותר.

8.2 הרעיון המרכזי של מבחן אנובה

הרעיון המרכזי של מבחן אנובה הוא להשוות את הפיזור של הנקודות מכל קבוצה ולהסתכל האם הפיזור בתוך הקבוצה הוא יותר הדוק מפיזור הנקודות בין הקבוצות. כלומר: 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 שהממוצע שלהן שונה באופן משמעותי.

לדוגמה: עם הדוגמה על הגברים האסיאתים, הנקודות בתוך הקבוצה של הגברים הלבנים יותר דומות אחת לשנייה בפיזור לעומת הנקודות בין קבוצת הגברים הלבנים לשחורים.

בצורה גרפית: נרצה שהגאוסיאנים יהיו יותר צפופים אחד מהשני ופחות מעורבבים אחד בין השני - ואם זה יהיה נכון: נוכל לומר כי ישנה הפרדה לפי הקטגוריה שהקבוצות מגיעות מהן:



8.3 הגדרה פורמלית

הגדרה - מבחן אנובה למשתנה אחד: נסמן בμ_i את התוחלת של המשתנה הi. 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 המשתנה הנ"ל תהיה זהה בין הקבוצות. ההנחה האלטנרטיבית אומרת כי מספיק שקיים זוג אחד שהתוחלת בניהם שונה. (זו הנחה מאוד מחמירה! מספיק שזוג אחד יראה לנו שההנחה לא סבירה בשביל שנשלול את השערת האפס).

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

נתונות n דגימות מq קבוצות: $n = n_1 + ... + n_q$.

תחת הנחות היסוד שלנו, נוכל לרשום את הערך של כל נקודה במדגם כ$y_{i,j} = \mu_j + \epsilon_{i,j}$ באשר $\epsilon_{i,j} \in N(0, \sigma)$ (סטיית התקן אינה ידועה - היא קיימה, לא ידועה, בהמשך נחשב או נשערך אותה). מדוע נוכל? אנחנו רושמים כל נקודה במדגם כסכום של תוחלת של הקבוצה של האיבר הספציפי ועוד איזשהו רעש - סטייה סביב התוחלת של הקבוצה שלו.

למשל, עבור גברים אסיאתים: הגובה הממוצע של גברים אסיאתים+ סטייה ימינה או שמאלה.

תחת ההנחה כי המודל הנ"ל של המשוואה הוא נכון, ניתן לבצע כל מיני חישובים. למשל:

א. ממוצע הנתונים בתוך הקבוצה הוא: $\overline{y_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j}$ באשר $1 \le j \le q$

ב. הממוצע של כלל הנקודות הוא: $\overline{\overline{y}} = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_j} y_{i,j} = \frac{1}{n}\sum_{j=1}^{q} n_j\overline{y_j}$

במבחן אנובה אנחנו רוצים לבדוק את פיזור הנקודות סביב הממוצע. לכן, בתחילה נגדיר את **פיזור הנקודות סביב הממוצע הכללי**:

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

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

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

הסבר עבור $*$ (זה לא היה בהרצאה, אני הוספתי למען ההבנה):

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

ומן הצד השני:

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

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

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

כעת עלינו להסביר מדוע: $-2\overline{\overline{y}}\sum_{i=1}^{n_j} y_{i,j} = -2\overline{y_j}\sum_{i=1}^{n_j} y_{i,j} + 2n_j\overline{y_j}^2 - 2n_j\overline{y_j}\overline{\overline{y}}$ ואז נוכיח את השוויון.

אם כן, מתקיים $\sum_{i=1}^{n_j} y_{i,j} = n_j\overline{y_j}$ ולכן:

$$-2n_j\overline{y_j}^2 + 2n_j\overline{y_j}^2 - 2n_j\overline{y_j}\overline{\overline{y}} = -2n_j\overline{y_j}\overline{\overline{y}}$$

וכן $-2\overline{\overline{y}}\sum_{i=1}^{n_j} y_{i,j} = -2n_j\overline{y_j}\overline{\overline{y}}$ כנדרש.

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

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

ומכאן: סכום הריבועים הכללי נגזר משני דברים:
א. הפיזורים של הקבוצות סביב הממוצע של הקבוצה
ב. המרחק בין הממוצע של כל קבוצה מהממוצע הכללי (ימין).
נחלק את הסכום הנ"ל לשניים:

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

ונאמר כי:
הפיזור סביב הממוצע של כל קבוצה:

$$SS_{Residual} = \sum_{j=1}^{q}\sum_{i=1}^{n_q}(y_{i,j}-\overline{y_j})^2 = \sum_{j=1}^{q}[(n_j-1)s_j^2]$$

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

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

ובמילים אחרות:

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

פיזור הנקודות סביב הממוצע הכללי נובע מ:

1. פיזור הנקודות סביב הממוצע של כל קבוצה.
2. פיזור הנקודות של ממוצע כל קבוצה סביב הממוצע הכללי.

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

א. מקור השונות - בין הקבוצות: ישנם $q-1$ דרגות חופש. סכום הריבועים הוא $S_{Between}$ ממוצע סכום הריבועים יהיה $S^2_{BETWEEN}=\frac{SS_{between}}{q-1}$

ב. מקור השונות - בתוך הקבוצות $n-q$. סכום הריבועים הינו $SS_{Residual}$ וממוצע סכום הריבועים יהיה: $S^2_{Residual}=\frac{SS_{Residual}}{n-q}$

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

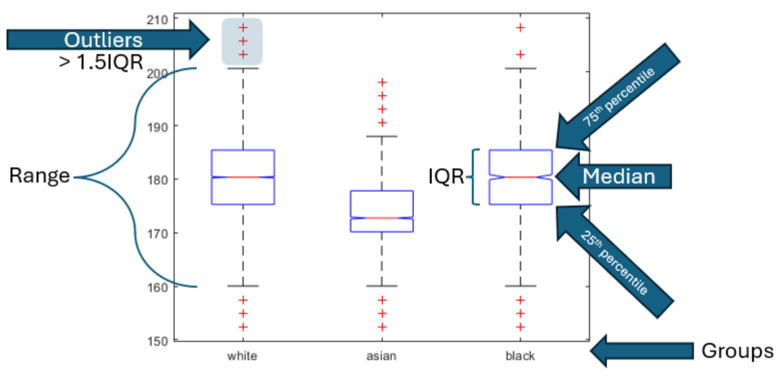
כפי שאמרנו, אנו רוצים לבחון את היחס בין פיזור הנקודות בתוך הקבוצות לבין פיזור הנקודות בין הקבוצות. לכן, המדד הסטטיסטי שנבחן הוא:

$$F=\frac{S^2_{Between}}{S^2_{Residual}}=\frac{\frac{SS_{between}}{q-1}}{\frac{SS_{Residual}}{n-q}}=\frac{\frac{\sum_{j=1}^{q}n_j(\overline{y_j}-\overline{\overline{y}})^2}{q-1}}{\frac{\sum_{j=1}^{q}[(n_j-1)s_j^2]}{n-q}}$$

ככל שF יותר גדול, אזי קל יותר לדחות את השערת האפס. מדוע? אם F גדל, אזי המונה גדל, כלומר $S^2_{Between}$ גדל והממוצעים רחוקים זה מזה - לכן הערך במונה גדל (ולכן השערת האפס לא נכונה) וכן בו זמנית, המונה קטן, כלומר $S^2_{Residual}$ קטן ולכן הגאוסיאנים נהיים צפופים (s_j קטן)

לאחר שחישבנו את F, הולכים לטבלה של התפלגות F עם $n-q, q-1$ ורמת המובהקות שרלוונטית לנו נמצא שם ערך x. אם $F>x$ הערך עבר את הגבול: נדחה את השערת האפס, ואחרת לא ניתן לדחות את השערת האפס.

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



הקו האדום מציין את החציון, הקופסה את הטווח של $25-75\%$ זה יקרא הIQR. הטווח יהיה $1.5IQR$ (קונבציה). מדוע זה עוזר? במקום להסתכל על ההתפלגות ולומר כמה הם שונים - נוכל להסתכל על דיאגרמה זו ולהסתכל על הנתונים. **לאחר ששללנו את השערת האפס: אנחנו יודעים שלפחות יש זוג אחד ששונה. עבור הדוגמה שלנו אם נריץ אנובה נקבל שאכן השערת האפס לא נכונה ממש - ואם נשתמש בדיאגרמת הקופסה, נוכל לראות שהזוגות שנראים שונים הם האסיאתים והלבנים.**

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

8.4 כיצד מתגברים על ההנחות?

8.4.1 איך מתגברים על ההנחה שהנתונים מתפלגים נורמלית?

מבצעים מבחן $Kruskal\ wallis$ (מבחן $Rank\ Sum$). מה עשינו בעבר כשלא יכולנו להניח דבר על הקלט - עברנו למבחנים אפרמטריים. הסתכלנו על הסדר וזה מה שיקרה כאן:

א. נחליף את הערכים בסדר שלהם במדגם.

ב. הסדר הממוצע בתוך קבוצה הוא $\overline{r_j} = \frac{1}{n_j}\sum_{i=1}^{n_j} r_{j,i}$ (יחליף את הממוצע של המשתנה בתוך הקבוצה)

ג.הסדר הממוצע של כלל המדגם הוא $\overline{\overline{r}} = \frac{1}{n}(1+2+......+n) = \frac{n+1}{2}$ (שהרי r מציין את הסדר בקבוצה והסדרים הם $1,...,n$.

ד. נתבונן בגודל הסטטיסטי הבא:

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

(דומה לאנובה, לא מדוייק).

מדוע $SSR_{Residual}$ לא במכנה? כיוון שלא נרצה להניח דבר על ההתפלגות בתוך הקבוצות.

אם יש לפחות 5 דגימות בכל קבוצה, אזי נוכל לומר כי $H \sim \chi^2_{q-1}$ (H מתפלג כמו χ^2 עם $q-1$ דרגות חופש).

אם אין אף ערך שחוזר על עצמו (**ולכן כל אחד מקבל מקום ייחודי משלו בדירוג הסדר**) אזי מתקיים $SSR_{TOTAL} = \frac{(n-1)n(n+1)}{12}$ ולכן במקרה זה:

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

8.5 כיצד נדע איזו קבוצה שונה?

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

בעברית: השוואות אנליטיות. באנגלית - $Post\ hoc\ tests$.

מה נוכל לעשות? נוכל לבצע בדיקת השערות בין כל זוג. כלומר, נעבור על כל הזוגות האפשריים. נזכר - כי אנחנו כעת עושים בדיקות מרובות - ולכן עלינו לתקנן את ה$P-value$ שלנו !

מדוע שלא פשוט נשווה בין הממוצע של כל הזוגות של כל הקבוצות? נעלה את הסיכון לשגיאה מסוג ראשון, אז למה שלא נתקנן? שכן זה יצור בעיות אחרות.

ישנה שיטה טובה יותר: $Post\ hoc\ test$ דואג למנוע שגיאות שכאלו.

ישנם סוגי מבחנים רבים תחת השם $Post\ hoc\ tests$ - התיקון של בונופרי הוא אחד מהם. מה ששונה ביניהם הוא בהנחות היסוד שלהן על הקלט והנתונים.

נדון במבחן שכזה בשם HSD:

א. נשווה כל זוג של ממוצעים כלומר $\mu_i - \mu_j$ לכל i,j.

ב. הנחות יסוד:

1. הנתונים בלתי תלויים זה בזה
2. הנתונים בכל קבוצה מתפלגים נורמלית
3. השונות בין הקבוצות דומה.

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

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

מהי התפלגות q? דוגמים n דגימות מk אוכלוסיות בעלות התפלגות זהה $N(\mu, \sigma)$. נסמן x_{min} כממוצע הקטן ביותר של אחת מן הקבוצות ואת x_{max} כממוצע הגדול ביותר של אחת מן הקבוצות, s^2 סטיית התקן של כלל הנקודות. ואז הוא מגדיר $q = \frac{x_{max} - x_{min}}{\frac{s}{\sqrt{n}}}$ שמתפלג בהתפלגות q. הרעיון הוא להסתכל על כל זוגות הדגימות שיש בעולם" ולבדוק היכן זוג הדגימות הנ"ל נופל בהתפלגות.

8.5.1 הפרש הממוצעים המינימלי שהוא מובהק סטטיסטית

ראינו כי $q_s = \frac{|\mu_A - \mu_B|}{SE}$. ולכן, $HSD = |\mu_A - \mu_B| = q_s SE = q\sqrt{\frac{S^2}{n}}$ (שכן $SE^2 = \frac{s^2}{n}$ ולכן $SE = \sqrt{\frac{s^2}{n}}$ באשר s^2 הוא האמד שלנו לשונות האוכלוסיה).

HSD הוא ההבדל המינימלי בין זוג ממוצעים שיחשב מובהק סטטיסטית (באותו סף בו חושב q) - כלומר HSD יהיה הסף שבו אנחנו נגיד - אם ההפרש בניהם יותר מזה, יש הבדל ביניהם שגרם לפער באנובה.

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 במדגם (30 ומעלה דרגות חופש) מתקיים כי q מתפלג בקירוב כ$q \approx \sqrt{2}z$ ולכן לרוב נלך לטבלת Z. ואז,

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

באשר n הוא מס' הדגימות בכל קבוצה.

8.5.3 סיכום HSD

א. נחשב $ANOVA$ - נרשום לפנינו את הפיזור בין הקבוצות ואת מס' דרגות החופש.

ב. נחשב את הערך הקריטי של q. - על סמך מס' הקבוצות, מס' דרגות החופש באשר N הוא גודל המדגם וk מס' הקבוצות מס' דרגות החופש יהיה $N - k$ ורמת המובהקות הדרושה.

ג. נחשב את ההפרש המינימלי שיחשב משמעותי סטטיסטית לפי $q\sqrt{\frac{SE^2}{n}}$

ד. נשווה כל זוג הפרשים ונבדוק אם הוא מעל הסף או שלא.

דוגמה:

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

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

לפי הטבלה המתאימה: q=3.63

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

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

8.6 אנובה בשתי משתנים

אנחנו מדברים כעת על מצב של: גברים מול נשים. האם יש הבדל משמעותי סטטיסטית בגובה של גברים לעומת נשים וגם לעומת גזע. כעת: כל קבוצה היא לפי 2 פרמטרים: גזע, ומין.

דוגמה נוספת: השוואה בין מין וקבוצת גיל (לא גיל עצמו כי זה משתנה רציף).

נוכל לרשום:

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

כל נקודה במדגם מוגדרת ע"י הממוצע הכללי של האוכלוסיה 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) ו$\Delta_{j,k}$ - שונות סביב הממוצעים.

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

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

כעת השערת האפס היא שלפי שני הפקטורים לא יהיה שוני בממוצעים. השערת האפס כאן היא מחמירה. ולכן מספיק שהשערה אחת תופר בכדי שנדחה את השערת האפס בכללותה.

כעת, מדובר בדיוק באותו פיתוח שעשינו עבור אנובה. נסתכל על הפיזור בתוך הקבוצות מול הפיזור של הקבוצות אחת יחסית לשנייה.

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

כעת, יש לנו שני מדדים סטטיסטיים ונעבוד בדיוק כמו קודם - נבדוק את שתי השערות האפס, אם שתיהן לא יתנו לדחייה - סה"כ השערת האפס כולה תתקבל. (וזה כמובן ניתן להכללה עבור יותר מ2 משתנים וכן הלאה).

הערה. יתכן שיש תלות גם בין המוצא למין. למשל: יש קשר בין גברים לאסיאתיים. לכן יהיה חישוב שונה - הדוגמה מופיעה במצגת של הרצאה 10 בעמוד 46.

הערה חשובה: אנובה הוא מבחן חד צדדי ימני.

הרעיון הוא שניתן לחשב את האנובה שלנו לפי שלושה דרכים שבהם העולם יכול להתנהג. (שכן מתוכנה סטטיסטית ומחישוב ידני התוצאות יכולות להיות שונות. מדוע? בדיוק בגלל הדיון הזה). זה נקרא $sum\ of\ squares$. הרעיון הוא - את מודל האנובה ניתן לחשב לפי כמה דרכים שהעולם יכול להתנהג: $1, 2, 3$.

סוג 1: יש למשתנים שלנו סדר מסויים. למשל: החלוקה הכי חשובה היא לפי הגזע, ורק אחר כך לאחר מין. או לחלופין. במקרה זה אנחנו מחשבים את SS לפי הסדר בו המשתנים מופיעים במודל. בסוג זה - מחשבים את הSOS לפי המשתנה הראשון, ואת המשתנה הבא מחשבים מהשארית של מה שנשאר לאחר הורדת הראשון. משתמשים כאשר יש סדר שרוצים להתחשב בו.

סוג 2: לא משנה הסדר המסויים, כל אחד יחושב בפני עצמו. מחשבים את הSOS בנפרד. משתמשים אם אין תלות.

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

סוג	1	2	3
תלוי בסדר המשתנים	כן. נחשב את ה- SS למשתנים לפי הסדר שבו הם מופיעים במודל	לא. כל SS מחושב בפני עצמו.	לא. מחשבים כל SS בפני עצמו + אינטראקציות
פיצוי למשתנים האחרים	רק למשתנים הקודמים בסדר המודל	כן	כן, וגם לאינטראקציות בין משתנים
מתי משתמשים?	משתנים בלתי תלויים, היררכיה, סדר בזמן	תלות חלשה או ללא תלות	תלות משמעותית או נתונים לא מאוזנים
מתי לא להשתמש?	יש מתאם בין גורמים		

מה עושים אם הנתונים לא מתפלגים נורמלית? כלומר - אנחנו ביותר ממשתנה אחד, האם יש מבחן דמוי קרוסקל? יש, הוא לא ממש מוצלח. לכן: לא נלמד אותה. ולכן מסקנה: אם הנתונים לא מפולגים נורמלית - אין לנו מבחן טוב עבור יותר ממשתנה אחד.

9 הרצאה 10: רגרסיה לינארית

מדובר במשפחה אחרונה של בדיקת אחרונות שנלמד, כעת אנחנו עומדים ללמוד כיצד לבדוק השערות במשתנים רציפים. עד כה: למדנו על בדיקת השערות במשתנה אחד ובקבוצות קטגוריאליות, לאחר מכן הרחבנו באנובה על בדיקת השערות עם משתנה אחד ויותר מ2 קבוצות (ואז גם הרחבנו ל2 משתנים..). כעת נדבר על בדיקת השערות במשתנים רציפים.

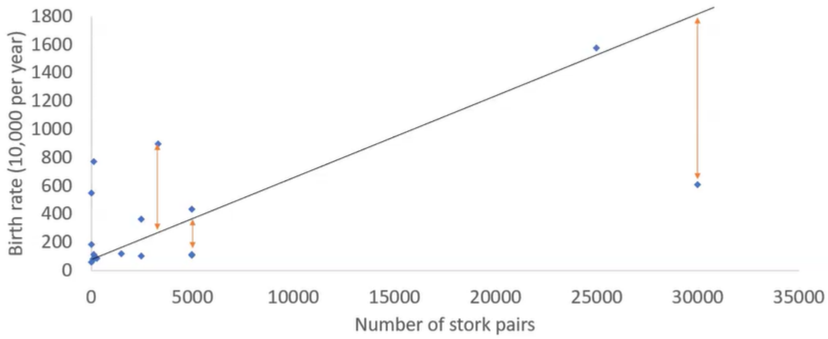
נתבונן בדוגמה ראשונה. השערה: חסידות מביאות ילדים לעולם. הרי, באירופה מגיעות מאפריקה לקראת האביב הרבה חסידות - ובאביב גם נולדים ילדים. אז, ברור שחסידות מביאות ילדים לעולם :)

זו השערה שקשה לבדיקה. מה אפשר לבדוק במקום? האם ישנו קשר בין חסידות לילדים? כלומר - האם ככל שיש יותר חסידות בעולם, ישנם יותר ילדים בעולם?

איך ניגש לבדוק השערה שכזו? ראשית, כמובן שנאסוף נתונים. נתבונן במס' החסידות במדינה מסויימת (משתנה רציף עד כדי עיגול) ובמס' התינוקות באותה מדינה (גם משתנה רציף). ההפיתוזה

שלנו - יש בניהם קשר. יש בניהם תלות (נניח לינארית). נקח את הנקודות, ונסמנם בגרף. נבחין כי אנחנו מעוניינים להעביר קו -קו מגמה.

איזה קו נעביר? **קו שסכום הריבועים ממנו יהיה מינימלי** (כלומר, המרחקים בערך מוחלט (בריבוע מן הסתם)) מן הקו. [ישנם קריטריונים ערכים שניתן להחליט, אנחנו נחליט על קריטריון זה].



אנחנו כמובן, נרצה לחשב את הקו הזה. נפרמל:

9.1 הגדרה פורמלית - מציאת קו המגמה

נניח כי נתון לנו מדגם ובו זוגות של נקודות (x_i, y_i) שאותן אנחנו רוצים לקרב על ידי ישר.

X מכונים (הציר האופקי): משתנים בלתי תלויים, משתנים מסבירים או מאפיינים.

Y מכונים (הציר האנכי): משתנה תלוי, $targets$.

משוואת הישר הינה $y = wx + b$ (באשר לעיתים נסמן $y = wx + b + \varepsilon$ באשר ε מסמל רעש)

הערך **על הישר** בנקודה x_0 הוא $(x_0, \hat{y_0})$

מכאן, ש**המרחק הריבועי של נקודה על הישר מהנקודה המתאימה במדגם** הינו: $\varepsilon_i = (y_i - \hat{y_i})^2$

כיצד מוצאים את משוואת הישר לפי הקריטריון הנ"ל?

$$E = min\{\varepsilon\} = min\{\sum_i \varepsilon_i\} = min\{\sum_i (y_i - \hat{y_i})^2\} = min_{w,b}\{\sum_i (y_i - w \times x_i - b)^2$$

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

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

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

מהמשוואה הראשונה נקבל כי $\sum y_i x_i = w\sum_i x_i^2 + b\sum x_i$.

מהמשוואה השניה נקבל $\sum y_i = w\sum x_i + nb$

נסדר את המשוואה השניה ונקבל: $b = \frac{1}{n}\sum y_i - \frac{1}{n}w\sum x_i \Longrightarrow b = \overline{y} - w\overline{x}$

נחזור למשוואה הראשונה,

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

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

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

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

ונקבל כי:

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

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

כעת, אנחנו יודעים בהינתן נתונים למצוא את b, w עבור קו המגמה שמבצע מינימום לסכום מרחק הריבועים.

כעת, נרצה לראות שיטות נוספות למציאת w, b. שיטות אלו מביאות לאותה תוצאה, אך ילמדו אותנו עוד.

9.2 שיטה שנייה (השיטה ההסתברותית)

נוכל להסתכל על הבעיה בראיה הסתברותית. כלומר:

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

משפט: אם X וY הם משתנים גאוסיינים בעלי התפלגות משותפת $f(x,y)$ ו$Variance$ סופי, פתרון ריבועים מינימליים הוא קו רגרסיה הנתון ע"י:

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

באשר $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$

ובמילים אחרות: $w = \frac{\sigma_{xy}}{\sigma_x^2}$

9.3 שיטה שלישית (אלגברה לינארית)

בשיטה זו נשתמש כאשר יהיה יותר ממשתנה אחד. באלגברה לינארית זה הולך להיות הרבה יותר פשוט.

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

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

וכן נתבונן בוקטור עמודה בגודל $n \times 1$ בשם $Y = \begin{pmatrix} y_1 \\ y_2 \\ .. \\ y_{n-1} \\ y_n \end{pmatrix}$. באופן כללי אם יש לנו $m-1$ משתנים אזי X היא מטריצה מגודל $n \times m$ 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 ש$Y = XW$. באופן כללי: Y הוא וקטור עמודה באורך n, X היא מטריצה בגודל $n \times m$ וW הוא וקטור עמודה בגודל m (כמס' המשתנים הבלתי תלויים + 1). הרעיון יהיה: בכפל מטריצות אחד - לסיים את הכל: ולא להכנס להרבה חישובים.

בשביל למצוא את W, נצטרך למזער את :

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

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

נגזרת של מטריצה

נגדיר:

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

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

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

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

מדוע? $z^TAz = \sum_{i=1}^n \sum_{j=1}^n z_i A_{ij} z_j$. ולכן, $\frac{\vartheta(z^TAz)}{\vartheta z_k} = \sum_{i=1}^n z_i A_{ik} + \sum_{j=1}^n A_{kj} z_j$. כיוון ש$A$ סימטרית, נקבל שזה שווה ערך לבדיוק $2\sum_{i=1}^n z_i A_{ij} = 2Az$.

כעת, נחזור למזעור ערך E שלנו:

$$E = (Y-XW)^T(Y-XW) = (Y^T-W^TX^T)(Y-XW) = Y^TY-Y^TXW-W^TX^TY+W^TX^TXW$$

$$Y^TY - 2W^TX^TY + W^TX^TXW$$

שכן המעבר מהשורה הראשונה לשנייה, התאפשר בזכות העובדה ש$Y^TXW = W^TX^TY$ כי מדובר בסקלר ולכן שמפעילים טרנספוז זה לא משתנה.

נגזור את הפונקציה שלנו, ונשווה לאפס:

$$\frac{dE}{dW} = -2X^TT + 2X^TXW = 0$$

$$X^TXW = X^TY$$

(הגזירה התבצעה לפי הכלל קודם, שהרי X^TX סימטרית ולכן הביטוי W^TX^TXW יגזר להיות $2X^TXW$ לפי הכלל הזה - $\frac{\vartheta}{\vartheta z}(z^TAz) = 2Az$).

כלומר המינימום של הפונקציה הינו: $X^TXW = X^TY$. כיצד נבודד את W? נכפול בשני הצדדים ב$(X^TX)^{-1}$

$$W = (X^TX)^{-1}X^TY$$

למשוואה זו יש שם: $Pseudo\ inverse$. נבחין כי המימד של X^TX הוא $m \times m$ - ולכן הגודל שלה: די קטן. [נעיר כי אי אפשר לפתוח לפי חוקי מטריצות, שכן X אינה מטריצה ריבועית אך X^TX כן כזו].

נעיר, כי אם $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ אזי $A^{-1} = \frac{1}{|A|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

מכאן, המשוואה הנ"ל נותנת לנו נוסחה סגורה לחישוב קו הרגרסיה W.

מדוע כדאי להשתמש באלגברה לינארית בשביל לפתור את מודל הרגרסיה (ולא במשוואה מהמודל הראשון?) שכן: בכל מצב שיש יותר ממשתנה בלתי תלוי אחד, המשוואות רק מתחילות להסתבך. ולכן: תמיד (!) להשתמש בשיטה זו.

לבסוף, לאחר שחישבנו את W נקבל את קו הרגרסיה.

זהו, יש קשר לינארי? נניח וקיבלנו את הקו $y = 0.0287 + 233.76$ - זו התאמה טובה? כיצד נדע לקבוע מהו טוב ומה לא טוב?

9.4 האם הקשר לינארי?

קיבלנו את קו הרגרסיה. מה הלאה?

נזכר במקדם המתאם של פירסון ־ שמקיים $-1 \leq r \leq 1$. אם $r = 1, -1$ הקשר מושלם (לינארי), אם $r > 0$ ישנו קשר לינארי חיובי בין המשתנים ואם $r < 0$ ישנו קשר לינארי שלילי.

נבחין כי לפי הנוסחה:

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

נתבונן בr^2:

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

שכן המעברים לפי מה שראינו כי $w = \frac{\sigma_{xy}}{\sigma_x^2}$.

כעת: SS_{REG} הוא סכום הריבועים שמסבירים כמה שונות נשאר בנתונים לאחר שהעברנו את קו הרגרסיה. במכנה, SS_{TOT} הוא סכום הריבועים על הנתונים המקוריים (כלומר כמה הע y סוטים מהממוצע שלהם). לכן: r^2 הוא אחוז ה$variance$ שמוסבר ע"י הרגרסיה.

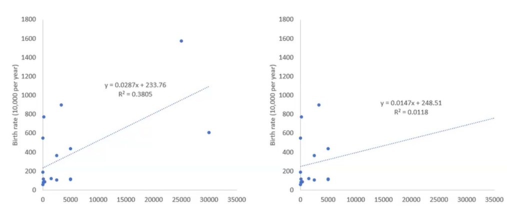
אם קיבלנו כי $r^2 = 0$ אזי $\hat{y}$ תמיד שווה לע y לא משנה מה נעשה. זה לא טוב! לעומת זאת: אם $r^2 = 1$ זה אומר כי $SS_{REG} = SS_{TOT}$ שכן כל פעם שנלך לאיזושהי נקודה במדגם, הע$\hat{y} - \overline{y}$ בריבוע שווה ממש למה שיש במדגם המקורי. ולכן: הקירוב ממש מצויין.

מינוח: אם $r^2 = 0$ הרגרסיה לא מסבירה כלום. אם $r^2 = 1$ הרגרסיה מסבירה הכל באופן מושלם.

כעת, אם נחזור לדוגמה הקודמת: $y = 0.0287 + 233.76$ חישבנו את $r^2 = 0.38$. מה זה אומר? זה אומר ש38% מהשונות מוסברים ע"י קו הרגרסיה. האם זה הרבה? האם זה מעט? זה תלוי ־ בכמה אנחנו מעוניינים להסביר. יתכן ואלעד יגיד ש38% מספיק לו בשביל להגיד: יש קשר בין החסידות לילדים. מישהו אחר יגיד לו ־ לא הסברת 62% אחרים מהשונות.

חשוב לשים לב! הרגרסיה רגישה מאוד לנתונים חריגים. אם יש לנו שני נתונים חריגים מאוד מהשאר ־ הם מאוד מאוד ישפיעו לנו על r^2.

בדוגמה הבאה: לאחר הסרת 2 הנקודות מ38% מהנתונים ירדנו להסבר על 1% בלבד:



9.5 חישוב מובהקות סטטיסטית

מה היינו עושים אם היה לנו יותר ממשתנה אחד? במקרה זה אי אפשר לצייר גרף עם שיפוע כי ישנו יותר ממשתנה אחד. לא נוכל לומר שישנה קורלציה. מה נעשה? נרצה למצוא דרך לעשות בדיקת השערות ־ **ולבדוק האם השיפוע של כל משתנה ב"ת מכלל המשתנים הב"ת שיש לנו שונה מאפס בצורה משמעותית או שלא?**

נכתוב השערות:

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

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

כיצד נעשה זאת? נסתכל על הסטטיסט $\frac{\hat{w_i}-0}{SE(w_i)}$.

הוריאנס של השארית ε מחושב ע"י $\sigma^2 = \frac{||y-XW||^2}{n-k} = \frac{RSS}{n-k}$ (כיוון שהראש ב"ת בכל המדידות ניתן לכתוב כי $\varepsilon \sim N(0, \sigma^2 I)$. כלומר זהו המרחק הריבועי בין y לשערוך 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 וk מס' המשתנים הבלתי תלויים (כולל הקבוע). כלומר, מסתכלים על ערך זה בהתפלגות t ועושים מבחן t עם $n-k-1$ דרגות חופש, וכרגיל: ברמת מובהקות מקבלים או שדוחים.

10 הרצאה 12: רגרסיה מתקדמת

10.1 רגרסיה לא לינארית

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 המטריצה גם את w_1 וגם את w_2.

10.2 רגרסיה של סדר $(Rank)$

ברגרסיה לינארית רגילה אנחנו מניחים כי הקשר קו ישר והנתונים מתפלגים נורמלית, אך מה אם יש ערכים חריגים או שהם אינם מתפלגים נורמלית? מקרה זה הוא מקרה שאם חשוב רק הסדר של הנקודות במדגם, ולא הערך שלהן (או אם יש ערכים שמפרים את הנחת הנורמליות). כלומר, אם למשל הנתונים הגיעו לנו לפי מיקומים בתחרות: דני הגיע ראשון, נועם הגיע שביעי וכן הלאה.

במקום להריץ רגרסיה על המספרים המקוריים (x_i, y_i) אנחנו הופכים כל מספר לדרגה שלו במדגם:

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

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

r הנ"ל שקול למתאם פרסון. הוא מתאם לפי סדר.

10.3 תיקון R^2 לגודל המדגם

ככל שנוסיף יותר משתנים ב"ת למודל, הוא יטה להתאים יותר לנתונים ולכן R^2 יגדל. (ככל שיהיו יותר משתנים, יהיה יותר רעש, אנחנו יותר נקרב למרות שזה לא באמת. לכן יש לנרמל את R^2 כנגד העובדה שמוסיפים עוד משתנים ב"ת).

לכן נתקנן את R^2 . נקבל כי:

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

באשר n מס' הנקודות במדגם וp הוא מס' המשתנים הב"ת.

מהי המשמעות של $\overline{R}^2$? נבחין כי הוא יכול להיות קטן מאפס עקרונית, וכן תמיד $\overline{R}^2 \leq R^2$ בהכרח. $\overline{R}^2$ יעלה רק אם העליה בR^2 גדולה מזו הצפויה באופן אקראי. כלומר: אם המשתנה החדש סתם אקראי $\overline{R}^2$ ירד ואם הוא באמת מסביר משהו $\overline{R}^2$ יעלה.

10.4 $Identifiability$

אם $X \in \mathbb{R}^{n \times p}$, עד כה הנחנו כי $rank(X) = p$. (לכן ניתן היה לחשב את $(X^TX)^{-1}$ - שכן X^TX מסדר $p \times p$. בשביל שהיא תהיה הפיכה $rank(X) = p$. אם כן, אם $p > n$, כלומר ישנם יותר משתנים מתצפיות, $rank(X) \leq n < p$ ולכן היא בהכרח לא תהיה p).

מה קורה כאשר $p >> n$? במקרה כזה:

1. w היא $not\ Identifiability$
2. אפשר לבצע חיזוי (יש פתרון מבין האנסופים הנתונים), אך המודל חסר ערך (כי יש אנסוף פתרונות אחרים).

מדוע ש$p > n$? כמה אפשרויות יתכנו:

1. קורלציה בין המשתנים
2. משתנים זהים מופיעים במודל בשמות שונים. (טמפרטורה במעלות ובפרנהייט).
3. יש פחות נתונים ממשתנים ב"ת. (למשל אם לוקחים בדיקה גנטית, יש 4 מליון עמודות - מקטעים של DNA ומס' השורות הוא האנשים עליהם יש לנו את הדגימה - מדובר על כמה מאות. כמובן שמס' המשתנים גדול מאוד ובמקרה זה $p > n$).

במצב זה לא נוכל להחליט איזה מהמשתנים בקורלציה עם המשתנה התלוי.

אם נרצה להעריך כמה המצב חמור (מבחינת $identifiablity$) אפשר להשתמש ב: $VarianceInflationFactor$.

VIF היא הדרך לחשב עד כמה הוריאנס של מקדם הרגרסיה גדול מערכו האמיתי כתוצאה מ $collinearity$.

10.5 $Variance\ Inflation\ Factor$

כשנרצה לבדוק כמה המשתנים שלנו מדברים זה עם זה, במקום לדבר עם המשתנה התלוי y, נשתמש בVIF. כלומר: כאשר השונות של מקדם הרגרסיה מתנפח בגלל מתאם גבוה בין שני משתנים מסבירים. **מאיפה הבעיה מגיעה? כשיש מתאם גבוה בין שני משתנים, קשה למודל לדעת להחליט למי מהם לתת את הקרדיט על השינוי בy.**

נניח כי אנחנו מחשבים מודל רגרסיה מהצורה:

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

נבצע את השלבים הבאים:

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

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

האינטואיציה: אם $R_i = 0$ כלומר לא הצלחנו להסביר אותו באמצעות האחרים, זה המצב הטוב לנו, בו $VIF = 1$ ואין לנו בעיה.

10.6 $Ridge\ Regression$

אם $rank(X) < p$ אזי אפשר להכניס אילוץ על פונקציית המטרה של הרגרסיה כדי לגרום למטריצה X^TX להיות הפיכה. בגרסה זו של רגרסיה, נמזער את השגיאה הריבועית תחת האילוץ $W^TW = c$ באשר c פרמטר. בעזרת כופלי לג'ראנז נוכל לכתוב את הפונקציה כ:

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

כאשר λ כופל לג'ראנז. אם נגזור נקבל:

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

למעשה כעת כיוון שנוספה מטריצה יחידה עם פקטור λ, היא כעת מטריצה ריבועית הפיכה.

10.7 מה עושים אם חלק מהנחות היסוד של רגרסיה לא מתקיימות?

הנחות היסוד של רגרסיה לינארית: חיזוי למשתנים תלויים רציפים, וריאנס קבוע לכל רמה של המשתנים הבלתי תלויים ($homoscedasticity$), השגיאה מתפלגת נורמלית.

בעולם האמיתי לפעמים קורים דברים אחרים: משתנה תלוי קטגוריאלי (חולה או בריא?), משתנה תלוי של ספירה (כמה ביקורים אצל הרופא?), התפלגויות לא סימטריות.

הפתרון הוא GLM. מרחיבים את הרגרסיה הלינארית בשתי צורות:
1. פונקציה מקשרת.
2. ההתפלגות של המשתנה התלוי היא ממשפחה אקספוננציאלית כלשהו.

לGLM שלושה רכיבים:
1. המשתנה התלוי שמגיע ממשפחה אקספוננציאלית. כלומר, אפשר לבטא את פונקצית הפילוג כאקספוננט. לדוגמה: נורמלית, פואסון, בינומי
2. חזאי לינארי: $\eta = \beta_0 + \beta_1 X_1 + .. + \beta_n X_n$
3. פונקציה מקשרת $g(\mu) = \eta$

מדוע צריך פונקציה מקשרת? אם המשתנה התלוי אינו רציף, לא נוכל לבנות מודל באופן ישיר כלומר $Y = X\beta$.

דוגמה לפונקציה מקשרת: עבור רגרסיה לינארית הפונקציה הינה פונקציית היחידה $g(\mu) = \mu$. עבור רגרסיה לוגיסטית שבה $y \in (0,1)$ נבחר $g(\mu) = log(\frac{1}{1-\mu})$. עבור רגרסיה פואסיאנית בה $y \in (0, \infty)$ נקח $g(\mu) = log(\mu)$.

הרצאה 13: $AB\ TESTING$