# שיטות סטטיסטיות ־ סיכום הרצאות למבחן

4 בדצמבר 2025

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 $6-4$) $\Leftarrow$ בדיקת השערות (הרצאה $9-7$) $\Leftarrow$ 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

ןם

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 $40-50$ 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<br>v | שכיחות<br>f(v) | שכיחות יחסית<br>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<br>v | שכיחות<br>f(v) | שכיחות יחסית<br>rf(v) | שכיחות יחסית מצטברת<br>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 $I=x_k-x_0$

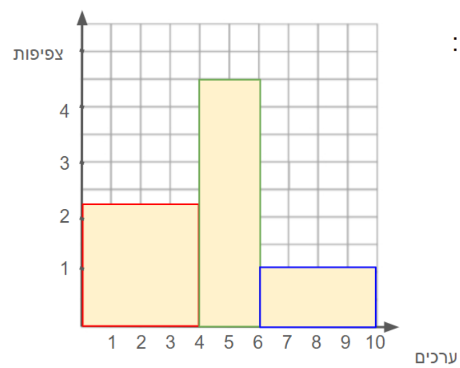
**מחלקה פתוחה:** רק גבול עליון או תחתון

| מחלקה | שכיחות f | רוחב מחלקה l | מרכז טווח | צפיפות d=f/l |
|---|---|---|---|---|
| 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 $min - u$

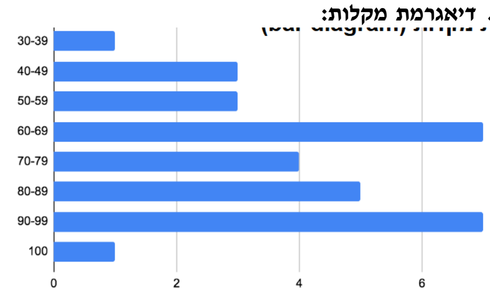
היסטוגרמה נכונה רק כאשר נתונים לנו כל הנתונים. אם נתון לנו טבלת שכיחויות אי אפשר לעשות זאת.

* תצוגה גרפית:
מייצגים נתונים באמצעות דיאגרמה.

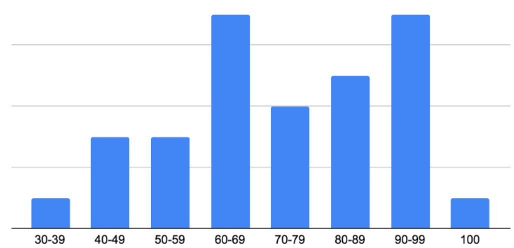
א. **דיאגרמת גבעול־עלה:** דיאגרמה בה מפצלים את הערכים לעשרות ויחידות. חלוקת טווח הערכים לגבעולים. וכן פירוט ערכי העלים.

| 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 |

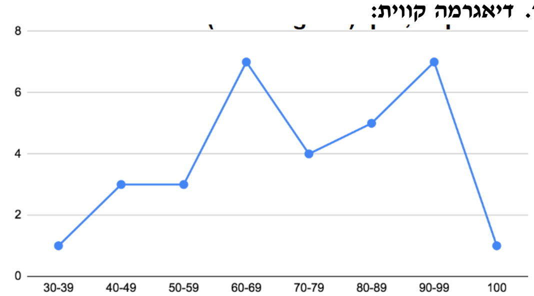
ב. דיאגרמת מקלות:



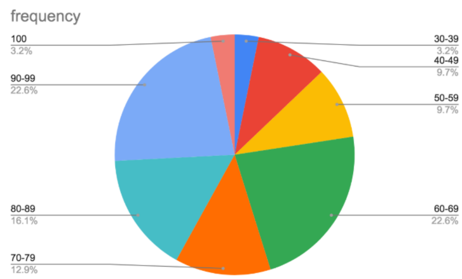

ג. דיאגרמת עמודות:




ד. דיאגרמה קווית:




ה. דיאגרמת עוגה:



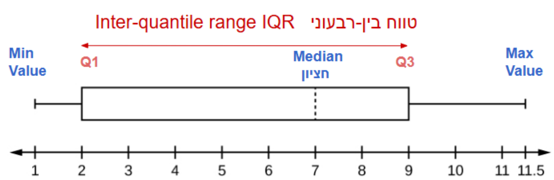

**מתי נשתמש באיזו דיאגרמה?**

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) ו$Q_2$ 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 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 ההפסש 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 הפונקצית 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.

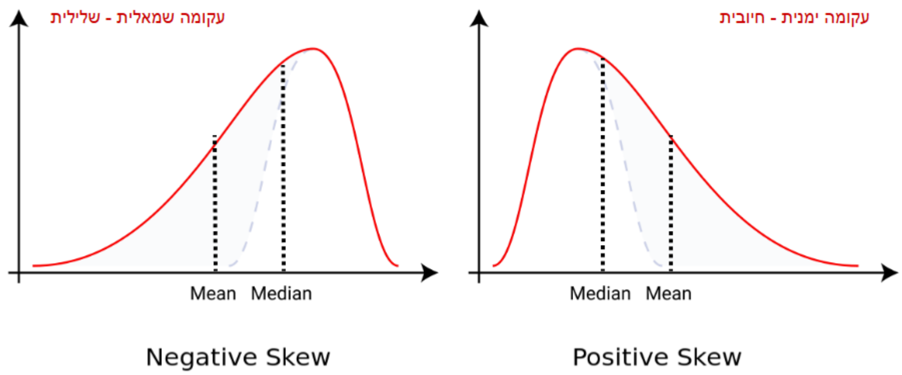
**ד. סכום ריבועיי השגיאות:** מרחקים ריבועיים של כל הערכים מהמדד $\sum_i (x_i - \overline{x})^2$
הממוצע מפחית למינימום את סכום ריבועי השגיאות.

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 ב$n$? 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 $x$ ל$y$.

נוכל להשתמש בדיאגרמת פיזור: על ציר ה$x$ ערכי ה$x$ ועל ציר ה$y$ ערכי ה$y$. 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 $\forall_{1 \leq i \leq N} X_i \sim F$.

**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 ש$\mu$ היא התוחלת ו$\sigma$ היא סטיית התקן. ($\sigma^2$ 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 $\mu$ 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 $\mu$ וסטיית תקן $\sigma$ 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 $\mu$ ושונות $\sigma^2$. כלומר לכל $1\leq i\leq n$ מתקיים $E[X_i]=\mu$. כך ש$S_n=X_1+X_2+...+X_n$.
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 $\forall_{1 \le i \le n} x_i \sim F$

**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 $\theta$ נסמן את האמד במדגם $\hat{\theta}$.

**דוגמה:** עבור התוחלת $\mu$ 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 $\mu$ 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 $\theta$ 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 ב$\mu$.

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 $\theta$ 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 $\theta$ 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, $Bias(\hat{\theta},\theta)=0\Longrightarrow E[\hat{\theta}]=\theta$. 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 $\mu$ 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 $\mu$ ושונות $\sigma^2$
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 $\theta_1$ ו$\theta_2$ הם אמדים חסרי הטיה עבור $\theta$ אזי נעדיף את זה עם השונות הקטנה יותר. את $\theta_2$ 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 $\theta_1$ ו$\theta_2$ הם אמדים שאינם חסרי הטיה עבור $\theta$ נעדיף את האומד $\theta_2$ 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 ה$k$ באמצעות ממוצע החזקה ה$k$ של התצפיות.

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

נראה כי $\mu_1$ הוא מרכז הנתונים, $\mu_2$ הוא דומה ומזכיר את השונות (הפיזור), $\mu_3$ מעיד על לאיזה כיוון העקומה הולכת, $\mu_4$ 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 ב$k$ 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 $\lambda$).

**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 ל$a$ ו$b$ הנדרשים.

**יתרונות השיטה:** קלה לחישוב, נוחה, ניתנה לחישוב עבור כל צורת התפלגות.
**חסרונות השיטה:** אם יש הרבה פרמטרים זה נהיה לא קל לחישוב, 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 ל$ln$? 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 $\theta$ כמכפלת ההסתברויות $x_1, ..., x_n \sim F$ בהינתן $\theta$:

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

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 $\theta$ 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 $\theta$ 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) | $= \bar{X}$ | כן |
| גאומטרי | X ~ G(p) | $= 1/\bar{X}$ | לא |
| מעריכי | X ~ Exp($\theta$) | עבור $\theta$: $= 1/\bar{X}$<br>עבור $\mu$: $= 1/\bar{X}$ | לא<br>כן |
| נורמלית:<br>- תוחלת<br>- שונות | X ~ N($\mu,\sigma^2$) | $= \bar{X}$<br>$= \sum_i (X_i - \bar{X})^2 / n$ | כן<br>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 $\mu$. 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 $\mu$ ושונות $\sigma^2$ 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 $\mu$ 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:

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

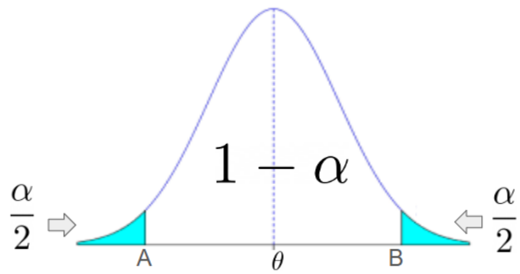
ישנה שקילות בגלל הערך המוחלט. המשמעות לשינוי זה היא - במדגם שגודלו $n = 25$ עבור מ"מ המתפלג נורמלית עם סטיית תקן 4 ותוחלת $\mu$ לא ידועה, בהסתברות 0.95 הרווח יכיל את $\mu$. כלומר - מצאנו אינפורמציה חשובה באשר ל$\mu$. נקרא לביטוי זה: **רווח הסמך של $\mu$ ברמה של 0.95.**

## 6.2 רווח סמך - הגדרה פורמלית

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

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

כך זה יראה בגרף:




הטעות האפשרית - היא $\alpha$ (הטורקיז) ורמת הסמך היא החלק הלבן.

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

עבור $X$ בעל תוחלת $\mu$, שונות $\sigma^2$ וגודל מדגם $n \geq 30$, וכן עבור $X$ נורמלי תוחלת $\mu$, שונות $\sigma^2$ 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 $\mu$ וסטיית תקן 22. נדגמו רנדומית 16 נורות ונמצא שאורך החיים הממוצע הוא 863. מצא רווח בסמך 90% ל$\mu$.

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 $\sigma^2$ 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 $n\geq 30$.

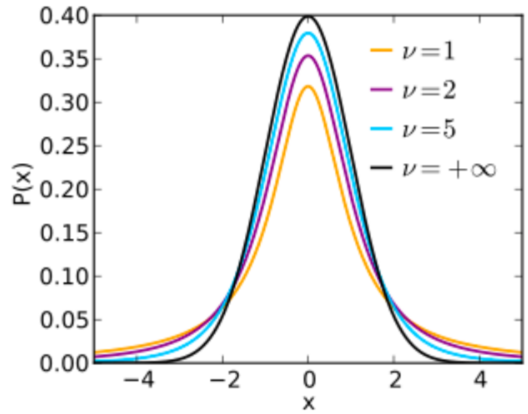
לכן, רווח הסמך ברמת סמך (רמת בטחון) של $1-\alpha$ אחוז עבור $\mu$ כאשר השונות אינה ידועה עבור מדגמים $n\geq 30$ היא:

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

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 ל$Z$ (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 $\alpha$ 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 $\alpha$ 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 $\alpha$ 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 $\overline{X} - ME < \mu < \overline{X} + ME$ .

# הרצאה 7: הסקת סטטיסטית - בדיקת השערות 1

# הרצאה 8: הסקת סטטיסטית - בדיקת השערות 2

# הרצאה 9: הסקה סטטיסטית

# הרצאה 10: רגרסיה

# הרצאה 11: $ANOVA$

# הרצאה 12: $AB\ TESTING$

# הרצאה 13: חזרה למבחן