סיכום בסטטיסטיקה – גל אור

מה זה בכלל בדיקת השערות?

היכולת להסיק מה קורה באוכלוסייה הכללית על סמך מדגם הרבה יותר קטן. לכן ניתן לאסוף מדגם קטן, לבצע עליו מחקר ומהמסקנות שיתקבלו להסיק מה קורה בכלל האוכלוסייה ממנה נלקח המדגם.

מה עושים בבדיקת השערות?

1. נגדיר מראש מהי האוכלוסייה הרלוונטית למחקר שלנו (למשל י כללית- מעל גיל 18, אוכלוסייה ספציפית- חולי איידס מעל גיל 40). המסקנות שנקבל יהיו רלוונטיות עבור האוכלוסייה שהגדרנו בלבד.
2. מהאוכלוסייה שהגדרנו ניקח קבוצה של אנשים שתייצג מדגם קטן מכלל האוכלוסייה. המדגם צריך היות קטן, מקרי ומייצג(ככל שהמדגם גדול הוא יותר טוב אך תמיד יחשב כקטן ביחס לאוכלוסייה).
3. נבצע ניתוח סטטיסטי על המדגם שלנו שממנו נסיק מסקנות.
4. לאחר קבלת המסקנות מהניתוח הסטטיסטי על המדגם נבצע תהליך של הכללה של המסקנות לאוכלוסייה כולה. כלומר נאמר שמה שהתקבל עבור המדגם הקטן כנראה רלוונטי גם עבור כלל האוכלוסייה שממנה נלקח המדגם ברמת ביטחון מסוימת (לא בוודאות).

שלב א: ניסוח השערות

תהליך בדיקת ההשערות מתחיל בניסוח שתי השערות מנוגדות (H מהמילה היפותזה = השערה) :

H0 – המצב הקיים הידוע באוכלוסייה (הנחת המוצא). שאנחנו מתחילים לחקור אנו חייבים להניח שהמצב הקיים לא עולה בקנה אחד עם מה שאנחנו חושבים שיקרה.(בד"כ מצביעה על היעדר קשר)

H1 – השערת החוקר , ההשערה המחדשת אותה אנו רוצים לאשש. השערה זו תמיד תהיה מנוגדת להשערת H0.

בפועל אנו ננסח את השערת החוקר בלבד, כלומר את השערת H1 , כלומר אילו קשרים אני מצפים לקבל במחקר שלנו. "מאחורי הקלעים" קיימת השערת H0 שהיא ההשערה המנוגדת להשערת H1.

ההשערה חייבת לכלול את המרכיבים הבאים:

1. לפחות שני משתנים.
2. ציון הקשר המצופה, כלומר את כיוון הקשר שאנו מצפים למצוא.
3. להיזהר מניסוח סיבתי (א' גורם לב').

סוגי השערות:

1. **השערה חד צדדית –** בעלת כיוון ברור של אופי הקשר בין משתנים.
   - יימצא קשר **חיובי** בין X לY.
   - יימצא קשר **שלילי** בין X לY.
2. **השערה דו צדדית –** ללא כיוון ברור של אופי הקשר בין המשתנים. (השערה כללית שקשה להגיד עליה משהו חד משמעי)
   - יימצא קשר **כלשהו** בין X לY.

איך מכריעים בין שתי השערות?

כדי להכריע בין H0 לבין H1 עורכים **מבחן סטטיסטי**.

התוצאה של המבחן הסטטיסטי הינה מהי ההסתברות שהנתון שחישבנו מהמדגם ייצג את (ישתייך ל) השערת H0 ? כלומר מכיוון שהשערת ה-0 היא המוצא, אני לוקחים את הנתון שחישבנו מהמדגם ושואלים איפה אותו נתון נופל על פני ההתפלגות של השערת ה-0 (מהי ההסתברות שאותו נתון ייפול על H0).

קוראים להסתברות זו "מובהקות המבחן" או p-value בקיצור Pv.ככל שההסתברות הזו נמוכה יותר, המשמעות היא שהמדגם הזה פחות שייך ל- H0 ויותר שייך ל- H1.

נרצה לקבוע את הגבול Xc (תחום חלקי) שיבחין לנו בין מצב בו השערת ה-0 נדחית או נלקחת.

אם ממוצע המדגם $\bar{X}$ נופל בתחום כלומר $\bar{X} > X_c$ דוחים את השערת H0 והשערת H1 נלקחת.

Xc - האיקס שאם ממוצע המדגם $\bar{X}$ יהיה גבוהה ממנו אז נוכל לדחות את H0 ולקבל את H1.

מבחן Z – חד צדדי

- משתנה רציף.
- כאשר יש כיוון להשערה חדשה (אותה אנו בודקים) ביחס להשערת המקור (אותה אנו רוצים לדחות).
- כאשר נתונה לנו סטיית התקן באוכלוסייה σ וממוצע המדגם $\bar{X}$.

דרך א':

1. נחשב את ערכו של Xc ע"י הנוסחה:

$$X_c = \mu_0 \pm Z_{1-\alpha} * \frac{\sigma}{\sqrt{n}}$$

μ_0 – התוחלת\הממוצע לפי השערת ה-0.

α – רמת מובהקות, הסיכוי שנדחה את השערת ה-0.

σ – סטיית התקן באוכלוסייה.

n – גודל המדגם.

$\pm$ - בהתאם להשערה האם היא גדולה מ- או קטנה מ-.

2. **אם מתקיים $X_c < \bar{X}$ דוחים את השערת H0 והשערת H1 נלקחת.**

דרך ב':

1. נחשב את ערכו של P_v (= השטח מ $\bar{X}$ עד לסוף הגרף) ע"י הפונקציה בשפת R : pnorm .
2. **אם מתקיים $P_v < \alpha$ דוחים את השערת H0 והשערת H1 נלקחת.**

מבחן Z – דו צדדי

שינוי בדרך חישוב הערך של Xc :

$$X_c = \mu_0 \pm Z_{1-\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

לבסוף בודקים אם התוחלת נמצאת בטווח של הרווח סמך השערת H0 נלקחת, אחרת השערת H1 נלקחת.

<u>מבחן T – חד צדדי</u>

- משתנה רציף
- כאשר נתונה לנו סטיית התקן במדגם S וממוצע המדגם $\bar{X}$. (כשלא נתונה סטיית התקן באוכלוסייה!)
- הבדיקה תלויה ברמות החופש $df = n - 1$.

דרך א':

1. נחשב את ערכו של Xc ע"י הנוסחה ובעזרת טבלת t :

$$X_c = \mu_0 \pm t_{1-\alpha,df} * \frac{s}{\sqrt{n}}$$

μ_0 – התוחלת\הממוצע לפי השערת ה-0.

α – רמת מובהקות, הסיכוי שנדחה את השערת ה-0.

df – רמות החופש.

S – סטיית התקן במדגם.

n – גודל המדגם.

$\pm$ - בהתאם להשערה האם היא גדולה מ- או קטנה מ-.

2. אם מתקיים $\bar{X} > X_c$ דוחים את השערת H0 והשערת H1 נלקחת.

דרך ב':

1. נחשב בעזרת טבלת t את הערך של $t_{1-\alpha,df}$.
2. נחשב את הערך של $t_{statistic}$ ע"י הנוסחה :

$$t_{statistic} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

3. נבדוק :
במקרה החיובי - האם מתקיים $t_{statistic} > t$ דוחים את השערת H0 והשערת H1 נלקחת.
במקרה השלילי - האם מתקיים $t_{statistic} < t$ דוחים את השערת H0 והשערת H1 נלקחת.

<u>מבחן T – דו צדדי</u>

שינוי בדרך חישוב הערך של Xc :

$$X_c = \mu_0 \pm t_{1-\frac{\alpha}{2},df} * \frac{s}{\sqrt{n}}$$

לבסוף בודקים אם התוחלת נמצאת בטווח של הרווח סמך השערת H0 נלקחת, אחרת השערת H1 נלקחת.

מבחן חי בריבוע – Chi Squared Test

לבדיקת טיב התאמה(תיאוריה מול מדגם)

- מבחן זה בא לבדוק האם אוכלוסייה מסוימת מתפלגת לפי התפלגות נתונה. המשתנה הנחקר מחולק למספר קטגוריות ויש לבדוק האם תוצאות המדגם תואמות להתפלגות הנתונה.
- משתנה לא רציף, לא נורמלי.
- כאשר המשתנה הוא קטגורי (קבוצות כדורגל, סוגי דם וכו').
- בהינתן התפלגות קטגורית תיאורטית של כלל האוכלוסייה ובנוסף התפלגות קטגורית של מדגם כלשהו, מטרת המודל לבדוק כמה שונה התפלגות המדגם מהתיאוריה.
- **לבדיקת ההשערות במודל זה נסמן:**

  <u>*Expected*, כלומר הצפי בהתאם לתיאוריה על כלל אוכלוסייה:</u>

  $P_{expected,i}$ – ההסתברות באחוזים של כל קטגוריה *i* בתיאוריה (לפי השערת ה-0).

  M_i – מספר הדגימות שהינו מצפים לקבל בתיאוריה בקטגוריה *i* מחושב ע"י $n * P_{expected,i}$.

  <u>*Observed*, כלומר תוצאות בהתאם למדגם מתוך האוכלוסייה:</u>

  $P_{observed,i}$ – ההסתברות באחוזים של כל קטגוריה *i* במדגם.

  X_i – מספר הדגימות שקיבלנו בפועל בקטגוריה *i* .

  n – גודל המדגם.
- **השערות המבחן:**

  H0 : $m_i = p_i * n$, *כלומר המשתנה מתפלג לפי התיאוריה* $observed \sim expected$.

  H1 : $m_i \neq p_i * n$, *כלומר המשתנה מתפלג אחרת מהתיאוריה.*
- **כלל ההכרעה :**

דרך א':

1. הערך הקריטי נקבע על סמך התפלגות חי בריבוע. התפלגות זו היא אסימטרית חיובית ותלויה בדרגות החופש $df = k - 1$ (k – מספר הקטגוריות).
   במידה ויש יותר מ2 סוגי הסתברויות אז נחשב
   $df = (\text{number of groups} - 1)\text{X}(\text{number of alternative distributions} - 1)$.
2. נחשב את ערכו של x^2 ע"י הנוסחה :

$$x^2 = \sum_{i=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{{x_i}^2}{m_i} - n$$

3. נחשב את ערכו של x_c^2 ע"י טבלת x^2 :

$$x_c^2 = x^2(df, 1 - \alpha)$$

   או לחילופין ע"י תוכנת *R* והפונקציה הבאה:

$$x_c^2 = qchisq(df, 1 - \alpha)$$

4. לבסוף בודקים : **אם מתקיים $x^2 > x_c^2$ השערת *H0* נדחית והשערת *H1* נלקחת.**

דרך ב':

1. נחשב את הערך של P_v (השטח מתחת לגרף) בתוכנת *R* :

$$P_v = pchisq(x^2, df, tail = false)$$

2. לבסוף בודקים:
   אם מתקיים $P_v < \alpha$ השערת *H0* נדחית והשערת *H1* נלקחת.

n	P Expected	Mi (counts expected)	P Observed	Xi (counts observed)	x^2
Category 1					
...					
Category i					
...					
Category K					
SUM	100%	n	100%	n	x^2

מבחן קולמוגורוב- סמירנוב

K-S Test

- מבחן זה הוא לא פרמטרי, כלומר אינו מחייב אותנו שההתפלגויות יהיו נורמליות.
- במבחן זה נוכל להשוות בין התפלגויות של משתנים רציפים או נומריים (נשים לב שגם משתנים רציפים יכולים להיחשב כמשתנים קטגוריים ע"י חלוקת ערכיהם לבינום).
- מבחן זה הינו מבחן המשווה בין התפלגויות ובודק עד כמה הן דומות. אפשר להשוות בין שתי התפלגויות של שני מדגמים, או בין מדגם והתפלגות תיאורטית. במידה ואנחנו רוצים לבחון נורמליות, נשווה את ההתפלגות שהתקבלה במדגם להתפלגות נורמלית-תיאורטית, כלומר הפיזור של ערכי המדגם בתנאי נורמליות.
- מבחן זה אינו קטגורי! הוא מתאים למשתנים רציפים או סדורים, כלומר שיש חשיבות לסדר (למשל שידוע a>b b>c c>d וכו').
- **השערות המבחן:**
  בהינתן 2 אוכלוסיות A ו-B נשאל:
  H0 : $A \sim B$, *כלומר* A *מתפלג כמו* B.
  H1 : $A \neq B$, *כלומר* A *אינו מתפלג כמו* B.
- **כלל ההכרעה :**
  הערך הקריטי שנקבע על סמך מבחן זה מחושב מתוך טבלת *K-S* ע"י ערכי רמת מובהקות וגודל המדגם.

$D_c(\alpha, n)$

α- רמת מובהקות.

n- גודל המדגם (במקרה שמדובר בהשוואת התפלגות מול תיאוריה).
במקרה ומדובר בשתי התפלגויות של מדגמים שונים A ו-B ננרמל את גודל המדגם ע"י החישוב :

$$n = \sqrt{\frac{A + B}{A * B}}$$

X	*counts*	*Prob_Observed*	*Cdf_Obs*	*Prob_Expected (dnorm)*	*Cdf_Exp (pnorm)*	*|Cdf_Obs- Cdf_Exp|*
Category 1						
...						
בעמודה זו תהיה חלוקה לקבוצות/קטגוריות	כמות הדגימות שנצפו בכל קטגוריה במדגם	ההסתברות של קטגוריה זו ביחס למדגם כולו (*Counts/n*)	הסתברות מצטברת- סכום העמודה משמאל עד לקטגוריה המתאימה	ההסתברות של ההתפלגות התיאורטית אליה אנו משווים	הסתברות מצטברת של התיאוריה- סכום העמודה משמאל עד לקטגוריה המתאימה	מעמודה זו ניקח את הערך המקסימלי והוא ייצא את הערך D_{max}

D_{max} – Maximum difference between observed cdf and expected cdf.

**** נשים לב שטבלה זו בנויה בהתאמה למקרה בו אנו משווים התפלגות של מדגם כלשהו לתיאוריה.**

לאחר שמצאנו את הערך של D_{max} נבדוק אם מתקיים:

$D_{max} > D_c$ השערת H0 נדחית והשערת H1 נלקחת.

לחילופין ניתן לבדוק אם מתקיים:

$P_v > \alpha$ השערת H0 נדחית והשערת H1 נלקחת.

מבחן חי בריבוע – Chi Squared Test

להשוואת משתנה קטגורי בכמה אוכלוסיות שונות

- נניח ש X משתנה קטגורי עם m קטגוריות שונות $x \in \{x_1, x_2, x_3, \dots, x_m\}$. (מחלות בקטגוריות למשל לחץ דם, השמנת יתר , אנמיה)
- המטרה במבחן זה היא להשוות את ההתפלגות של X בין K אוכלוסיות שונות $A \in \{A_1, A_2, A_3, \dots, A_K\}$.
- **השערות המבחן:**
  H0 : $A_1 \sim A_2 \sim \dots \sim A_k$, *כלומר X מתפלג דומה בכל הקבוצות\האוכלוסיות.*
  H1 : *אחרת, כלומר X אינו מתפלג דומה בכל הקבוצות\האוכלוסיות.*
- נשים לב שבמבחן זה מדובר בהשוואה בין מספר כלשהו K של אוכלוסיות בעוד שבמבחן *k-s* מדובר בהשוואה של 2 התפלגויות בלבד. בנוסף מספר הקטגוריות במבחן *k-s* הוא הרבה יותר גדול כיוון ששם מחלקים משתנה שהוא רציף לקטגוריות. היחס בין הקטגוריות של X במבחן זה הוא שהן לא תלויות זו בזו בשונה מבמבחן *k-s*.
- נסמן:
  $O_{i,j}$ – מספר התצפיות\דגימות בקטגוריה i באוכלוסייה j.
  n_j – גודל המדגם באוכלוסייה j.
  t_i – מספר התצפיות\דגימות הכולל בקטגוריה i בכל האוכלוסיות.
  $P_{i,j}$ – ההסתברות לקבל את מקרה קטגוריה i באוכלוסייה j, מחושב ע"י: $\frac{O_{i,j}}{n_j}$
  P_i – ההסתברות הכוללת בקטגוריה i בכלל האוכלוסייה, מחושבת ע"י : $\frac{t_i}{n}$
  $E_{i,j}$ – מספר הדגימות המצופה באוכלוסייה j בקטגוריה i תחת השערת ה0, מחושב ע"י : $\frac{t_i}{n} * n_j$
- **שלבים:**
  1. נמלא את הטבלה הבאה על מספר התצפיות בכל מקרה לפי הנתונים:

Population / Category	A_1	A_2	…	A_j	…	A_k	Total (t_i)
x_1							t_1
x_2							t_2
…							…
x_i				$O_{i,j}$			t_i
…							…
x_m							t_m
Total (n_j)	n_1	n_2	…	n_j	…	n_k	n

טבלת Observed

  2. לאחר מילוי הטבלה על מספר התצפיות בכל מקרה נוכל לבנות טבלה חדשה שבה נציג את המידע של הטבלה הקודמת באחוזים של הסתברויות עבור כל מקרה:

$$P_{i,j} = \frac{O_{i,j}}{n_i}$$

Population / Category	A_1	A_2	…	A_j	…	A_k	Total (t_i)
x_1							t_1
x_2							t_2
…							…
x_i				$P_{i,j}$			t_i
…							…
x_m							t_m
Total (n_j)	n_1	n_2	…	n_j	…	n_k	n

3. על מנת להמשיך באנליזה נצטרך לבצע את מבחן חי בריבוע chi squared .
על מנת נוכל לבצע מבחן זה צריכות להיות לנו observed וגם expected. הטבלה שחישבנו בשלב 2 היא הobserved ועתה ונצטרך לחשב טבלה בעבור expected.
כלומר בטבלה זו יופיע כמה אני תצפיות אני מצפה שיהיו בכל קטגוריה בהינתן גודל המדגם של כל אוכלוסייה.

$$E_{i,j} = \frac{t_i}{n} * n_j$$

טבלת Expected

Population Category	A_1	A_2	...	A_j	...	A_k	Total (t_i)
x_1							t_1
x_2							t_2
...							...
x_i				$E_{i,j}$			t_i
...							...
x_m							t_m
Total (n_j)	n_1	n_2	...	n_j	...	n_k	n

בטבלה זו נשים לב:

אם ב- 80% מסך התאים בטבלה זו יש ערך $E_{i,j} \leq 5$ זה אומר שאין מספיק מופעים במבחן זה וצריך לאחד קטגוריות.

4. בשלב זה נחשב את x^2 ע"י הנוסחה :

$$x_{i,j}{}^2 = \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Population Category	A_1	A_2	...	A_j	...	A_k	Total
x_1							
x_2							
...							
x_i				$x_{i,j}{}^2$			
...							
x_m							...
Total $x_j{}^2$	$x_1{}^2$	$x_3{}^2$	...	$x_j{}^2$	...	$x_k{}^2$	X^2_{total}

דרך א - לאחר חישוב ערכו של X^2_{total} נותר לחשב את ערכו של X^2_c, נעשה זאת ע"י אחת מהדרכים:

נחשב את ערכו של x^2_c ע"י טבלת x^2 :

$$x^2_c = x^2(df, 1-\alpha)$$

או לחילופין ע"י תוכנת R והפונקציה הבאה:

$$x^2_c = qchisq(df, 1-\alpha)$$

לבסוף בודקים : **אם מתקיים $x^2_{total} > x^2_c$ השערת *H0* נדחית והשערת *H1* נלקחת.**

דרך ב' - במקום לשבת את X^2_c נוכל לחשב את נחשב את הערך של P_v בתוכנת R :

$$P_v = pchisq(x^2, df, tail = false)$$

3. לבסוף בודקים: **אם מתקיים $P_v < \alpha$ השערת *H0* נדחית והשערת *H1* נלקחת.**

<u>מבחן להשוואת תוחלת בין שתי קבוצות לא מצומדות ע"י מבחן T</u>

- קבוצות לא מצומדות – קבוצות בלתי תלויות, כלומר בכל קבוצה יש נבחנים שונים. אנו מניחים כי אין קשר בין המדגם הראשון לבין המדגם השני, כל מדגם הוא מדגם מקרי מתוך אוכלוסייה.
- מדגמים הם בלתי תלויים כאשר:
  - ← המשתתפים נדגמים באופן אקראי מאוכלוסיות שונות.
  - ← המשתפים נדגמו מאוכלוסייה אחת, אך הוקצו באופן אקראי לשתי קבוצות טיפול שונות.
- מטרת המבחן היא לבחון את השאלה האם התוחלות\הממוצעים של שני המדגמים שונים זה מזה. במבחן למדגמים בלתי תלויים, אנחנו רוצים להסיק על הפרש הפרמטרים $\mu_1 - \mu_2$ של התכונה הנחקרת.
- נשתמש במבחן זה כאשר:
  - ← $\sigma^2{}_2, \sigma^2{}_1$ שונויות באוכלוסיות אינן ידועות ושוות (נלמד בהמשך לבדוק שוויון שונויות).
  - ← המדגמים לא מצומדים
  - ← המשתנים בכל אוכלוסייה מתפלגים נורמלית
- למרות שאיננו יודעים את $\sigma^2{}_2, \sigma^2{}_1$ על מנת שנוכל לקדם את בחינת השערת ההבדלים נפעל בהנחה שהשונויות באוכלוסיות זהות אחת לשנייה וגם לשונויות במדגמים(בהמשך נלמד כיצד לבדוק זאת) :
  - ← $\sigma^2{}_2 = \sigma^2{}_1 = S^2$
  - ← S^2 הוא הממוצע המשוקלל של השונויות במדגמים $s^2{}_2, s^2{}_1$.
- נסמן :
  - ← Y - משתנה נורמלי
  - ← $\overline{Y_2}, \overline{Y_1}$ - ממוצעי המדגמים
  - ← S_1, S_2 - סטיות התקן של המדגמים
  - ← n_1, n_2 – מדגמים בלתי תלויים
  - ← α – רמת המובהקות
- השערות המבחן :

H0 : $\mu_1 = \mu_2 \leftarrow \mu_1 - \mu_2 = 0$

H1: לגבי השערת המחקר ישנן כמה אפשרויות:

$$\mu_1 < \mu_2 \leftarrow \mu_1 - \mu_2 < 0$$
$$\mu_1 > \mu_2 \leftarrow \mu_1 - \mu_2 > 0$$
$$\mu_1 \neq \mu_2 \leftarrow \mu_1 - \mu_2 \neq 0$$

- חישוב שונות המדגמים המשוקללת:
ממנה נוכל לחשב את סטיית התקן (=שורש השונות)

$$\boxed{S^2 = \frac{(n_1 - 1)\cdot S_1^2 + (n_2 - 1)\cdot S_2^2}{(n_1 - 1) + (n_2 - 1)}}$$

- חישוב דרגות החופש :

$$\boxed{df = n_1 + n_2 - 2}$$

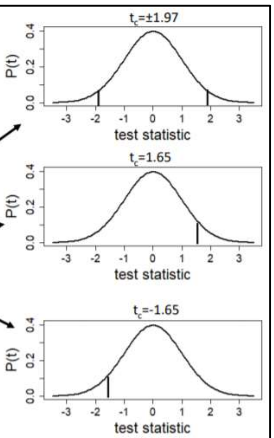
-
- חישוב ערכו של $t_{statistic}$:

$$\boxed{T = \frac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{\overline{Y_1} - \overline{Y_2}}{s\cdot\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}$$

- שלב ההכרעה:
נחשב את t_c לפי דרגות החופש ורמת המובהקות ונכריע את המבחן.
המספרים בגרפים רק להמחשת המיקום של t_c בגרפים.
במקרים אלו השערת H0 נדחית והשערת H1 נלקחת.

General Hypothesis Testing

Alternative Hypothesis	Rejection Region	R code
H_a: $\mu_1 \neq \mu_2$	$\|T\| > t_{1-\alpha/2,df}$	qt(1-α/2,df)
H_a: $\mu_1 > \mu_2$	$T > t_{1-\alpha,df}$	qt(1-α,df)
H_a: $\mu_1 < \mu_2$	$T < t_{\alpha,df}$	qt(α,df)



הבדלה בין מבחנים:

1. **מצומד או לא מצומד(אם גודל המדגמים שונה זה בטוח לא מצומד)**
2. **האם המידע נורמלי(נורמליות משותפת)**
3.