שיטות סטטיסטיות - הסיכום

מחבר - אופק אבן דנן

מרצים – רעות צרפתי, אלעד יום טוב | מערכי תרגול: שני פינקלשטיין

עדכני לפני המבחן

תוכן עניינים

סטטיסטיקה תיאורטית -

מושגים בסיסיים –

אוכלוסייה (population) –
אוסף של אנשים, דברים, אובייקטים אותו אנו רוצים ללמוד

מדגם (sample) –
תת-קבוצה (מייצגת) של האוכלוסייה, המדגם משמר את התכונות של כלל האוכלוסייה, את הפיזור וגם ניתן להכליל ממנה אל כל השאר.

דגימה (sampling) –
אופן בחירת המדגם, יש מספר אופנים לבחירת מדגם –

דגימה הסתברותית
- **דגימה רנדומית** דגימה רנדומית של k מתוך Nעם החזרה או ללא החזרה
- **דגימה בשכבות** חלוקת האוכלוסייה לשכבות זרות ומשלימות שבהן תהיה דגימה רנדומית לפי הגודל של כל שכבה.
- **דגימת** חלוקת כל האוכלוסייה לקבוצות זרות ומשלימות דגימה רנדומית של קבוצות והוספת כל הפרטים מכל קבוצה למדגם
- **דרגת חופש -** מספר הערכים החופשיים להשתנות בניתוח נתונים, בהתחשב במגבלות מסוימות (כמו חישוב ממוצע או מגבלות אחרות).

דגימה לא הסתברותית
- דגימת נוחות הכול בבת אחת – עבור פיילוטים
- דגימה שיפוטית לפי שיקול דעת החוקרת, לפי מענה על שאלונים – חקר על קבוצה קטנה.
- דגימת כדור-שלג "חבר מביא חבר" – תופעות נדירות

משתנה (variable) –
תכונה הניתנת לתצפית ולמדידה עבור כל אלמנט באוכלוסייה –

סוגי משתנים –
- **קטגורי -** קבוצת ערכים סופית למשל מידות, דרגות.
- **מספרי בדיד -** מספר אנשים.
- **מספרי רציף -** גובה, משקל מרחק.

סולמות מדידה –
- **סולם שמי (nominal) -** יחס זהות, ללא יחס סדר כמו קטגוריה מגדרי, ארץ לידה
- **סולם סדר (ordinal) -** יחס זהות, עם יחס סדר, תווית מידה , דרגה אקדמית
- **סולם רווחים (interval) -** עם יחס סדר, עם מרווחים קבועים כמו טמפרטורה
- **סולם מנה (ratio) -** יחס סדר, מרווחים קבועים, נקודת אפס כמו גובה ומשקל

ערך (datum) –
הערך שנמדד עבור משתנה יחיד באוכלוסייה

מידע (data) –
הערכים שנמדדו עבור כל האוכלוסייה במשתנה מסוים.

סטטיסטי (statistic) –
ערך המחושב על סמך הדאטא, **התצפיות בפועל**, כלומר על סמך כל הערכים שנמדדו.

פרמטר
תכונה של האוכלוסייה המקורית, למשל תוחלת או שונות.

הצגת המידע –
קיימים מספר דרכים להצגת המידע –

הצגה טבלאית

- **טבלת שכיחויות –** טבלה המציגה כמה אותו ערך מופיע בכלל המדגם.

שם $name$	שכיחות $f(name)$	שכיחות יחסית $F(name)=\frac{f(x)}{N}$	שכיחות יחסית מצטברת $COL=\sum_{names\ before} F(name)$
אופק	$f(Ofek)=2$	$\frac{2}{7}$	$\frac{2}{7}$
רואי	$f(Roi)=5$	$\frac{5}{7}$	1

- **שכיחות יחסית –** מה האחוז מכלל האוכלוסייה עם הערך הזה.
- **שכיחות יחסית מצטברת -** מה האחוז מכלל האוכלוסייה עם הערך הזה או פחות (מתקבל דרך הסכימה של כל הערכים שלפני).
- **חלוקה למחלקות –** ניתן לחלק משתנה מסוים לכמה קטגוריות, למשל אנשים בטווח גיל:

הצגה גרפית

- **היסטוגרפית צפיפויות –** מציג במלבנים על ציר ה-X המייצגים את הטווח של המחלקה ובציר ה-Y מוצגת השכיחות היחסית.

שם $X{:}(s,f)$	צפיפות מחלקה $Y{:}d=\frac{f(s,f)}{\Delta(s,f)}$	רוחב מחלקה $\Delta(s,f)=f-s$	שכיחות $f(s,f)$	שכיחות יחסית $F(name)=\frac{f(x)}{N}$
$(\mathbf{0},\mathbf{5})$	$\frac{10}{5}$	$5-0=5$	$f(0{,}5)=10$	$\frac{10}{30}$
$(\mathbf{5},\mathbf{10})$	$\frac{20}{5}$	$10-5=5$	$f(5{,}10)=20$	$\frac{20}{30}$

- **דיאגרמת גבעל –** מראים בגרף כמה מתוך האוכלוסייה נמצאים בטווח מסוים.
- **דיאגרמת עמודות / מקלות –** מראה כמה יש מכל קבוצה מסוימת מתוך המדגם.
- **דיאגרמת קווית –** מראה את העלייה או את הירידה – נוח להראות התקדמות.
- עוגה – מראה את החלק היחסי ובהינתן שאין היררכיה בין הקבוצות.

מדדים סטטיסטים

- **שכיח (mode) -** הערך עם השכיחות הגבוהה ביותר$(f(x_i))$ $\bar{x}=\arg\max$
- **מרכז הטווח (midrange) -** הממוצע בין התצפית הגבוהה והנמוכה ביותר $\bar{x}=\frac{1}{2}(x_1+x_n)$
- **חציון (median) -** 50% או פחות גבוהים ממנו, 50% או פחות נמוכים ממנו
  עבור n זוגי $\arg(k)$ $2k=n\rightarrow$, ל- n אי זוגי $k=\left\lfloor\frac{n}{2}\right\rfloor\rightarrow\frac{\arg(k)+\arg(k+1)}{2}$
- **ממוצע (mean) -** סכום כל הערכים מחולק במספר התצפיות. $\frac{1}{n}\sum_{i=1}^{n}\arg(i)$
- **מינימום / מקסימום –** הערך הנמוך/ הגבוהה ביותר והנמוך ביותר מכלל הערכים שנמדדו.
- **רבעון ראשון ושלישי –** החציון של החציון הראשון והחציון של החציון השני.
- **שונות (Variance) –** פיזור הנתונים ביחס לממוצע $\frac{\sum_{i=0}^{n}(x_i-\bar{x})^2}{n}$ $S_x^2=\sigma_x^2=$ סטיית תקן == שונות בריבוע

- **חישוב מדד מיקום –** משתנה רציף
מחלקה: m | גבולות: $L_1 - L_0$ | שכיחות: f | מצטברת: F | אחוז: k

$$Md = L_0 + \frac{\frac{n \cdot k}{100} - F(x_{m-1})}{f(x_m)}(L_1 - L_0)$$

Box Plot –

מציג את הנתונים בצורה חזותית המספקת מידע על הפיזור, על החציון ורמת החריגות. הוא מכיל תיבה שבה החלק התחתון מיצג את Q_1 , הגבול העליון של התיבה ייצג את Q_3, ובאמצע התיבה יהיה קו שייצג את החציון. מתוך התיבה יצאו קווים למעלה ולמטה כל עוד נמצאים בטווח של $1.5 \times$ מהטווח של ההפרש הרבעונים (בין הרבעון הראשון לשלישי). מעבר לטווח הזה הנתונים שעשויים להיות במקסימום ובמינימום והם מחוץ לטווחים יקראו Outliers.

קביעת מיקום מרכזי –

- **מספר השגיאות -** כמה מהערכים אינה שווים למדד עצמו $|\{x_i | x_i \neq \overline{x}\}|$
- **השגיאה המקסימלית -** המרחק המקסימלי מהמדד עצמו $\max\ |x_i - \overline{x}|$
- **סכום השגיאות המוחלטות -** מרחקים אבסולוטיים של כל הערכים מהמדד $\sum_i |x_i - \overline{x}|$
- **סכום ריבועי השגיאות -** מרחקים ריבועיים של כל הערכים מהמדד $\sum_i (x_i - \overline{x})^2$

	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

מדדי הפיזור המוגדרים ביחס למיקום המרכזי –

- **אחוז השגיאות -** אחוז התצפיות השונות ממדד מיקום מרכזי $\frac{1}{n}|\{i | x_i \neq \overline{x}\}|$
- **גודל השגיאה המקסימלית -** המרחק הגדול ביותר ממדד מיקום מרכזי (ממוצע או מרכז טווח).
- **הטווח הבין רבעוני -** הטווח בו נמצאים 50% הערכים המרכזיים בהתפלגות IQR .
- **ממוצע הסטיות המוחלטות -** ממוצע מרחקי התצפית ממדד מיקום מרכזי $\frac{1}{n}\sum_i |x_i - \overline{x}|$
- **ממוצע ריבועי הסטיות -** ממוצע ריבועי מרחקי התצפית ממדד מיקום מרכזי $\frac{1}{n}\sum_i (x_i - \overline{x})^2$

שונות וסטיית תקן –

- **באוכלוסייה –** $\sigma_x = \sqrt{\frac{1}{n}\sum_i (x_i - \overline{x})^2} \quad , \quad {\sigma_x}^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$
  במדגם - $S_x = \sqrt{\frac{1}{n-1}\sum_i (x_i - \overline{x})^2} \quad , \quad {S_x}^2 = \frac{1}{n-1}\sum_i (x_i - \overline{x})^2$ באופן כללי במדגם נתייחס ל- $n-1$.
- **עבור טבלת שכיחויות -** ${S_x}^2 = \frac{1}{n}\sum_i x^2 f(x) - \overline{x}^2$
- **חוק צ'ביצ'ב –** לפחות $1 - \frac{1}{k}$ מהערכים הם במרחק k סטיות תקן מהממוצע.

תקנון -

טרנספורמציה ליניארית מערכי ההתפלגות האמפירית $Z_x = \frac{x - \overline{x}}{S_x}$

קשר בין משתנים –

ניתן למדוד את הקשר בין משתנים באמצעות קשר ליניארי, באמצעות מדד המתאם של פירסון.
נראה בהמשך בפרק של רגרסיה, המדד בין $-1 \leq r \leq 1$ ככל שיותר רחוק מ- 0 יש יותר תלות.

$$r = \frac{\sum_{i=1}^{n} z_{xi} \cdot z_{yi}}{n} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{x})}{n \cdot S_x \cdot S_y} = \frac{cov(x,y)}{n \cdot S_x \cdot S_y}$$

הסקה סטטיסטית –

- **הסקה דדוקטיבית –** מהכלל לפרט
- **הסקה אינדוקטיבית –** מהפרט לכלל.

נוכל לכמת את מידת הוודאות שבאי הוודאות באמצעות הכלים בקורס.

מדגם מקרי –

מדגם מקרי בגודל n מתוך משתנה מקרי X הוא מדגם של n משתנים מקריים (תצפיות) כך ש-

א. $X_1 \dots X_n$ הם מ"מ בלתי תלויים.

ב. לכל משתנה מקרי X_i יש את אותה פונקציית הסתברות כמו של X. לכן, לכל i מתקיים - $X_i \sim F$.

משפט - שקילות מדגם מקרי –

דגימה מקרית עם החזרה של n איברים מתוך אוכלוסייה עם תכונה $X \sim F$ שקולה למדגם מקרי בגודל n מתוך מ"מ $X \sim F$ ולהיפך.

ולכן מבחינה מעשית הדגימה המקרית תתבצע מתוך האוכלוסייה כאשר נבצע דגימה של משתנה מקרי. ומצד שני, נוכל להשתמש במה שאנחנו יודעים על מ"מ על כל דגימה מקרית.

צורת ההתפלגות תלויה במספר גורמים:

- בסוג ההתפלגות באוכלוסייה.
- בסוג הסטטיסטי.
- בגודל המדגם.

התפלגות הדגימה – התפלגות הדגימה של סטטיסטי מסוים s עבור מדגמים בגודל n שנלקחו מאוכלוסייה בה ערכי המשתנה מתפלגים לפי התפלגות F.

משפט - תוחלת הסטטיסטי –

ממוצע המדגם המסומן $\overline{X}$ הוא מ"מ בעל פונקציית הסתברות שאפשר לחשב עבורו תוחלת ושונות.
תוחלת הסטטיסטי (ממוצע המדגם $\overline{X}$) שווה לתוחלת המ"מ X שממנו דוגמים $E[X] = E[\overline{X}]$.
באופן דומה ניתן לבצע חישוב השונות וסטיית התקן של $\overline{X}$. נקבל: לשונות - $\sigma^2(\bar{X}) = \frac{\sigma^2(X)}{n}$
ולסטיית התקן $\sigma(\bar{X}) = \frac{\sigma(X)}{\sqrt{n}}$.

הצדקה לשימוש במדגם אחד –

במדגמים שונים עבור אותה אוכלוסייה יש ממוצעים שונים, אם נדגום הרבה מדגמים ונחשב ממוצע אז הממוצע של כל הממוצעים יתקרב לממוצע של כלל האוכלוסייה. למה שנרצה לדגום הרבה מדגמים? מדגם אחד עשוי לתת לנו את מה שנרצה לכן השאלה המתבקשת היא מה הסבירות שבמדגם יש סטייה גבוהה מהמוצע / תוחלת של האוכלוסייה.

לכן, נתעניין במידת הפיזור של ההתפלגות של הדגימה של ממוצע המדגם. סטיית התקן של ממוצע המדגם תהיה סטיית התקן תהיה $\frac{\boldsymbol{\sigma(X)}}{\boldsymbol{\sqrt{n}}}$. לכן, ככל שהמדגם גדול יותר השונות של סטיית התקן של הממוצע תהיה קטנה יותר.

אי שוויון צ'ביצ'ב –
במקרה של המשתנה המקרי X – $P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$ במקרה של התפלגות הדגימה של הממוצע X: $P\left(\mu - k\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}$ כאשר k הוא המרחק במספר סטיות תקן.

החוק החלש של המספרים הגדולים –
כאשר n שואף לאינסוף אז – נבחר - $k \geq \frac{\epsilon\sqrt{n}}{\sigma}$ ועבור משתנה מקרי X עם תוחלת ושונות מתקיים:

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

כאשר $n \to \infty$ נקבע שהתוצאה תשאף ל- 1. כלומר ככל שהמדגם גדל הערך של ממוצע המדגם מתקרב לתוחלת המקורית.

הקרבה של הערך במדגם לערך של כלל האוכלוסייה-

משפט –
בדגימת מדגם שגודלו n מתוך מ"מ **המתפלג נורמלית** עם תוחלת μ וסטיית תקן σ יהיה ממוצע המדגם $\overline{X}$ מתפלג נורמלית גם הוא עם התוחלת μ וסטיית התקן $\frac{\sigma}{\sqrt{n}}$.

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

משפט הגבול המרכזי –
יהי X משתנה מקרי כלשהו בעל תוחלת μ וסטיית תקן σ, ויהיו $X_1 \dots X_n$ משתנים מקריים בלתי תלויים כאשר לכל אחד מהם יש התפלגות זהה לשל- X אז- כאשר $n \to \infty$ ההתפלגות של הממוצע של $\overline{X}$ שואפת להתפלגות הנורמלית עם התוחלת של μ וסטיית תקן $\frac{\sigma(X)}{\sqrt{n}}$.

כלומר כאשר המשתנה המקרי X לא נלקח מהתפלגות נורמלית, ההסתברות של הממוצע מדגם של X היא כמעט נורמלית (גדול מספיק הכוונה ל**מדגם גדול מ- 30**).

משפט הגבול המרכזי קובע שאם לוקחים אנחנו לוקחים מדגמים חוזרים בגודל מספיק קובע שאם לוקחים מדגמים חוזרים בגודל מספיק $n \geq 30$ אז התפלגות ממוצעי המדגם תהיה נורמלית גם אם האוכלוסייה תהיה נורמלית.

מציאה מהטבלה בהתפלגות נורמלית –

$$P(X \leq y) = P(Z_X \leq Z_y) = \Phi(Z_y)$$

כאשר X זה המשתנה המתפלג נורמלי, Z_x, Z_y אלו הערכים המתוקננים, ו- Φ מייצג את הערכים בטבלה, עבור $n > 30$ נחפש בטבלת Z, אחרת בטבלת t.

בעיית האמידה -

- **האמדים estimator -** סטטיסטי שבעזרתו אומדים פרמטר הסתברותי לא ידוע, למשל חישוב תוחלת לא ידועה למדגם באמצעות ממוצע.
- **האומדן estimate -** התוצאה שקיבלנו עבור האמד במדגם הספציפי.

נתון משתנה מקרי $X \sim F$ מדגם מקרי של $X_1 \dots X_n$ בלתי תלוי כאשר לכל i מתקיים $X_i \sim F$. לדוגמה, מהם ערכי הפרמטרים של פונקציית ההסתברות או הצפיפות $X \sim F$?

טרמינולוגיה

- **θ -** פרמטר באוכלוסייה
- **$\hat{\theta}$ -** פרמטר במדגם – אמד סטטיסטי.

לדוגמה – עבור התוחלת μ נבחר אמד, שזה הממוצע של המדגם $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

דגשים –

לאותו האמד ייתכנו תוצאות שונות למדגמים שונים באותה האוכלוסייה, לכן האמד הוא משתנה מקרי בעצמו ויש לו התפלגות דגימה. תכונותיו של האמד תהיה התפלגות **הדגימה** של הסטטיסטי שתלוי רק בערכי המדגם.

- **שגיאת האמידה estimation error –** המרחק בין ערך האמד לערך הפרמטר האמיתי $\hat{\theta} - \theta$.
- **הטיה של אמד estimator bias –** תוחלת שגיאת האמד - $Bias(\hat{\theta}, \theta) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta$.

תכונות של אמדים -

התכונות לקביעת מהו אמד טוב.

- **עקביות – consistent –** ככל שהמדגם גדל, ההסתברות שהאמד יתכנס לפרמטר האמיתי גדל ׳כלומר, ההפרש בין הפרמטר לאמד קטן $\hat{\theta} \xrightarrow[n \to \infty]{} \theta$.
- **חוסר ההטיה –** ההטיה של האמד שווה לאפס $Bias(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta = 0$.
- **אינווריאנטיות פונקציונלית –** אם T אנ"מ ו- g פונקציה חח"ע אזי אבור g: $g(T)$ הוא גם אנ"מ.

אמדים לדוגמה –

- **ממוצע -** $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ – אכן חסר הטיה, ועקבי.
- **שונות מתוחלת ידועה –** $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$ - עבור תוחלת תהיה חסרת הטיה.
- **שונות עבור מדגם לתוחלת לא ידועה -** $\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ חסר הטיה.

יעילות אמדים –

- אם θ_1, θ_2 אמדים חסרי הטיה נעדיף את θ_2 בעל השונות הקטנה יותר $Var(\theta_2) < Var(\theta_1)$.
- במקרה הכללי – הטעות תוגדר כתוחלת ריבועי השגיאות-

$MSE(\hat{\theta}, \theta) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var_{\theta}(\hat{\theta}) + Bias_{\theta}(\hat{\theta}, \theta)^2$, לכן, אם θ_1, θ_2 אמדים נעדיף את θ_2 בעל הטעות הקטנה יותר $MSE(\theta_2) < MSE(\theta_1)$, כמו כן זה חלק גם על המקרה הראשון.

בחירת אמדים –

- בדיקה האם $E[\theta] = E[X] = \mu$
- בדיקה למי יש את השונות הקטנה ביותר במידה והאמדים חסרי הטיה) באמצעות חישוב והשוואה בין שתי השונויות. אחרת, נחשב את ה- MSE שזה ממוצע שגיאות האמידה והשוואה ביניהם (בשלב זה ה- $Bias_{\theta}(\hat{\theta}, \theta)$ אמור להיות שווה ל-0).

שיטות אמידה מקובלות -

שיטת המומנטים

היא שיטת אמידה על פי פרמטרים המאפיינים התפלגות של אוכלוסייה מסוימת. עבור משתנה מקרי המתפלג F עבורה ישנם k פרמטרים בלתי ידועים, נגדיר פונקציית מומנטים, כל מומנט יאמוד באמצעות ממוצע החזקה ה- k של התצפית.

$\mu_1 = E[X^1] = g_1(\theta_1 \dots \theta_n)$ – תוחלת | $\mu_3 = E[X^3] = g_3(\theta_1 \dots \theta_n)$ – צידוד

$\mu_2 = E[X^2] = g_2(\theta_1 \dots \theta_n)$ – שונות | $\mu_5 = E[X^4] = g_4(\theta_1 \dots \theta_n)$ – גבנוניות

כאשר כל מונומנט במדגם בנוי כ- $\widehat{\mu_k} = \frac{\sum_{i=1}^{n} {X_i}^k}{n}$.

בפועל, משווים כל מומנט מסדר k לאומדן שלו במדגם ונפתור מערכת של k משוואות ל- k נעלמים.

היתרונות של השיטה הזו היא שהשיטה היא כללית, פשוטה לחישוב ונכונה עבור כל צורות ההתפלגויות. החסרונות הם שהשיטה מתאימה בעיקר עבור מעט פרמטרים ועלולים לקבל אמדים מוטים או לא סבירים.

שיטת הנראות המרבית –

שיטה זו מנסה להעריך את הפרמטרים של המודל כך שהנתונים הנצפים יהיו בעלי הסתברות מקסימלית להתרחש, כלומר, מחפשים את ערכי הפרמטרים שממקסמים את פונקציית הנראות.

באופן כללי, עבור משתנה מקרי X עם פרמטר לא ידוע θ, פונקציית הנראות מוגדרת כמכפלת פונקציות ההסתברות עבור כל התצפיות: $f(x;\theta) = \prod_{i=1}^{n} L(\theta)$ כך למשל עבור התפלגויות שבהן יש ערך מספרי כמו פואסון פונקציית הנראות תוגדר כמכפלה של פונקציות ההסתברות לכל ערך בתצפית. לעומת זאת, במקרים שבהם מודדים הצלחות נרצה לבחור נראות המבוססת על התצפית הסכום של מספר ההצלחות. ולכן נשתמש בפונקציה הרגילה עם משתנה p.

הערכים שמביאים את הנראות למקסימום הם הערכים הסבירים ביותר. אז בעבור כל ערך של p ניתן לחשב את פונקציית הנראות L לשם הנוחות לעיתים נעטוף הכול ב-ln, זה אפשרי מאחר שהפונקציה $\ln(\,)$ היא מונוטונית ואז נמצא ערך מינימום על ידי גזירה. למשל עבור בינום:

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

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

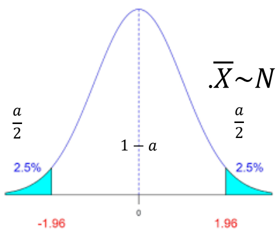
$$\Big(\ln\big(L(p|k,n)\big)\Big) = 0 = \frac{k}{p} - \frac{n-k}{1-p} \Rightarrow p = \frac{k}{n}$$

נראות עבור כל סוגי ההתפלגויות (מומלץ מאוד לצרף לדף נוסחאות)-

סוג המשתנה המקרי	פונקציית ההתפלגות	תוחלת	שונות	אמד נראות מרבית – אנ"מ	חסר-הטיה?
התפלגות ברנולי $X \sim Ber(p)$	$P(X=x) = \begin{cases} 1 & x=1 \\ 1-p & x=0 \end{cases}$	p	$p(1-p)$	$= \frac{X}{n}$	כן
התפלגות בינומית $X \sim Bin(n,p)$	$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$	$= \frac{X}{n}$	כן
התפלגות גיאומטרית $X \sim Geo(p)$	$P(X=k) = (1-p)^{k-1} \cdot p$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$= \frac{1}{\bar{X}}$	לא (כלפי מטה)
התפלגות אחידה בדידה $X \sim U(a,b)$	$P(X=k) = \frac{1}{b-a+1}, k \in (a, a+1 \dots b)$	$\frac{a+b}{2}$	$\frac{1-(b-a+1)^2}{12}$	$= max\{X1 \dots Xn\}$	לא (כלפי מטה)
התפלגות מעריכית $Exp(\lambda)$	$P(x,\lambda) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	עבור λ: $= \frac{1}{\bar{X}}$ עבור μ: $= \frac{1}{\bar{X}}$	לא (כלפי מעלה) כן
התפלגות פואסון $X \sim Pois(\lambda)$	$P(X=k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$	λ	λ	$= \bar{X}$	כן
נורמלית	$\frac{1}{\sqrt{2\pi\sigma^2}}^{-(\frac{(x-\mu)^2}{2\sigma^2})}$	μ	σ	תוחלת: $\bar{X} =$ שונות $= \frac{\sum_i (X_i - \bar{X})^2}{n}$	כן לא (כלפי מטה)

רווח סמך וטווח -t

- **אמידה נקודתית** – על מדגם מחושב סטטיסטי אחד משערך את הפרמטר.
- **רווח סמך לממוצע המדגם** – על סמך המדגם המחושב מטווח של ערכים. $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$.
- **רווח סמך באופן כללי** – הרווח $A.B$ הוא רווח סמך ברמה של $1-a$ עבור θ.
$P(A<\theta<B)=1-a$



רווח סמך עבור התפלגות נורמלית כאשר השונות ידועה –

- עבור x בעל תוחלת μ, שונות σ^2 וגודל מדגם $n\geq 30$ – מתקיים - $\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right)$
- עבור x בעל תוחלת μ, שונות σ^2 וגודל מדגם $n>0$ – מתקיים –

$$P\left(-z<Z_{\overline{X}}<z\right)=\Phi(z)-\Phi(-z)=2\Phi(z)-1$$

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

גודל מדגם	התפלגות	שונות	הטווח:
$n>30$	התפלגות דגימה נורמלית	ידועה	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(\mu,\frac{\sigma^2}{n})$
	התפלגות דגימה נורמלית	לא ידועה	$\frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}\sim N(\mu,\frac{\hat{S}}{n})$
$n\leq 30$	דגימה מאוכלוסייה מתפלגת נורמלית	ידועה	$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\sim N(\mu,\frac{\sigma^2}{n})$
	דגימה מאוכלוסייה מתפלגת נורמלית	לא ידועה	$\frac{\overline{X}-\mu}{\hat{S}/\sqrt{n}}\sim t(n-1)$

התאמה בהתפלגות t	התאמה בהתפלגות Z
$\Rightarrow \boldsymbol{P\left(\mu-t_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}<\overline{X}<\mu+t_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}\right)=1-a}$	$\Rightarrow \boldsymbol{P\left(\mu-Z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}<\overline{X}<\mu+Z_{\frac{a}{2}}\frac{\sigma}{\sqrt{n}}\right)=\underbrace{1-a}_{\text{רמת ביטחון}}}$

- **מרווח שגיאה** - $EBM=\left(Z_{\frac{a}{2}}\right)\cdot\left(\frac{\sigma}{\sqrt{n}}\right)=\epsilon=\frac{|mistake|}{2}$, ניתן להפוך ולקבל - $n=\left(\frac{Z\cdot\sigma}{EBM}\right)^2$

שימו לב – טבלת ה- t מציינת את השטח מימין לנקודה לכן אין צורך לעשות אחד פחות.

בדיקת השערות –

בחלק זה - ממוצע של האוכלוסייה = μ | ממוצע של המדגם = $E[X]$.

קבלת החלטה –

		ההחלטה	
		D_0	D_1
האמת	H_0	✓	שגיאה מסוג 1
	H_1	שגיאה מסוג 2	✓

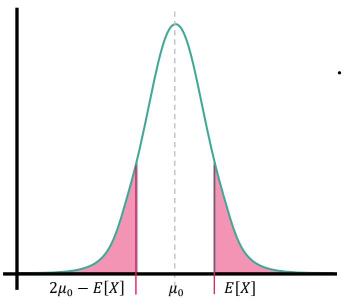
נגדיר –

- **הנחת האפס H_0-** - ההנחה הקיימת, מה שחושבים שהוא המצב.
- **H_1** - הנחה שטוענת שהנחת האפס איננה נכונה
- **שגיאה מסוג 1 – (דחיית H_0 בטעות) –** מקבלים את H_1 בטעות, (FP), סיכוי לטעות $= \alpha$.
- **שגיאה מסוג 2 – (מקבלים את H_0 בטעות) –** דוחים את H_1 בטעות, (FN), סיכוי לטעות $= \beta$.

לכן מרחב הביטחון הם - $1-\alpha$ לסוג 1 ו- $1-\beta$ לסוג 2. קיים בין α ו- β יחס הפוך.

חישוב הסבירות של השערת האפס –

נניח שבודקים את ממוצע מדגם האוכלוסייה $E[X]$ השונה מהממוצע של האוכלוסייה μ_0, נרצה שהמדד עבור הסבירות של השערת האפס יהיה קטן יותר ככל שההבדלים בין גדול יותר. יחושב כפונקציה של סטיית התקן מסביב לממוצע - $\frac{E(x)-\mu_0}{SD(x)/\sqrt{n}}$



מבחני סף –

P value הוא ההסתברות לקבל את התוצאה שנצפתה בהנחה שנכונה.

- **מבחן right-tailed –** יש חשד ש- $E[X] > \mu_0$.
- **מבחן left-tailed -** יש חשד ש- $E[X] < \mu_0$.
- **מבחן דו-צדדי two-tailed -** אין חשד לאף כיוון ספציפי.

פתרון באמצעות נקודה קריטית –

נגדיר את הערך הקריטי על פי סוג המבחן –

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

פתרון באמצעות Z –

1. חישוב Z_α
2. מתכננים את סטטיסטי המבחן. $Z_x = \frac{x-\bar{x}}{S_x}$
3. בודקים אם האי שוויון בין Z_α לסטטיסטי המתוקנן Z_x בדומה למבחן קריטי.

חישוב ה p value –

ה-p-value עוזר לנו להכריע אם יש לנו עדות מספקת לדחות את השערת האפס עבור H_1 במידה והמשתנים נורמליים המדד מחושב כך עבור מבחן Z – $P(\mu_0) = 2\left(1 - \Phi\underbrace{\left(\frac{|E(x)-\mu_0|}{SD(x)/\sqrt{n}}\right)}_{z}\right)$

זהו ה-Type 1 error. מדד זה נמצא בין 0 ל- 1 רציף וככל שהמרחק בין התוחלת מדגם לממוצע גדול יותר הוא קטן יותר. המשמעות היא שזה "הצדדים" של גרף הפעמון כאשר כל צד הוא המרחק בין התוחלת לבין ממוצע המדגם. הוא לא אומר אם לקבל או לדחות את השערת האפס.

קבלה/ דחיית השערת האפס –

1. להחליט את סף ה- p value (a) לפני החישוב (גדול קטן או אין).
2. לחשב את ה- p value של המדגם.
3. לדחות את H_0 אם ה- p value נמצא בצדדים.

נשים לב – זוהי איננה קביעה חד משמעית, זה אומר שאי אפשר לדחות או לקבל את השערת האפס, יכולות להיות סיבות כמו חוסר בנתונים, רעש וכו'.

בדרך כלל הסף המקובל הוא 0.05 כיוון שהוא מספיק להעביר את המסר אך לא דורש דיוק גדול. במשתנה נורמלי הסף הוא $Z = 1.96$. אם לא מצוין הסף הוא 0.05.

דרך אחרת לחישוב p value–

ההסתברות לקבל את המדגם שקיבלנו בהנחה שהשערת האפס היא הנכונה, אם החלטנו לדחות את השערת האפס כיוון שה-p-Value היה נמוך מהסף שקבענו, נאמר שהשערת האפס נדחתה באופן מובהק סטטיסטית.

תיקון Bonferroni–

במידה ונעשים מספיק מבחנים עבור סף p value הסבירות לקבל תוצאה קטנה מהסף עולה ככל שיש יותר מבחנים ולכן עולה הסיכוי לטעות מסוג 1. התיקון אומר יש לקבוע את הסף ל-p-value כ- $a_{Bonferroni} = \frac{\alpha}{k}$ כאשר α הוא הסף למבחן בודד ו-k הוא מספר המבחנים כלומר מספר הזוגות עליהם נבחן את ההשערות, ונבצע מבחנים בהתאם בין הזוגות.

תיקון Šidák –

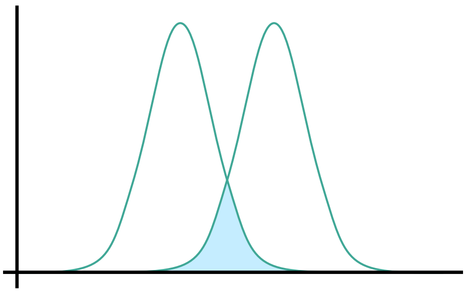
תיקון פחות מחמיר לתיקון רמת המובהקות כאשר מבוצעות בדיקות מרובות, נגדיר את a להיות - $a_{Šidák} = 1 - (1 - a)^{\frac{1}{n}}$ משמש להפחת הסיכוי לטעות טעות מסוג 1.

חישוב גודל מדגם דרוש –

$$P(\mu_0) = 2\left(1 - \Phi\left(\frac{|E(x) - \mu_0|}{\frac{SD(x)}{\sqrt{n}}}\right)\right) \Rightarrow n = \frac{SD^2(x)}{(E(x) - \mu_0)^2} z_\alpha^2$$

מדגמים מרובים –

כאשר יש מספר מדגמים צריך לקחת בחשבון את הפרמטרים של שתי ההתפלגויות ואת השכיחות היחסית שלהן, נרצה לראות את השטח המשותף ששתי ההתפלגויות יוצרות.

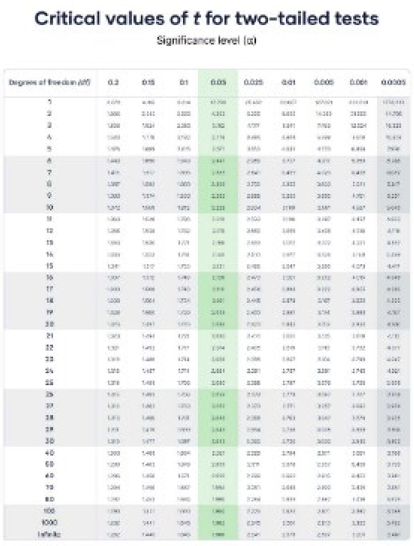


עוצמת המבחן

עוצמת המבחן המשלים לסיכוי לשגיאה מסוג 2 – כלומר הסיכוי לקבל את השערת האפס בטעות. העוצמה תלויה במבחן, ברמת המובהקות, בנתונים ובגודל הנתונים ובגודל האפקט. עוצמת המבחן כתלות בגודל המדגם (טבלה מצורפת) מראה שככל שהמדגם גדול יותר צריך הפרש קטן יותר בין הקבוצות.

גודל האפקט: ה-d של כהן (Cohen's d) -

מודגר – $d = \frac{\overline{x_1}-\overline{x_2}}{s_p}$ כאשר $s_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$



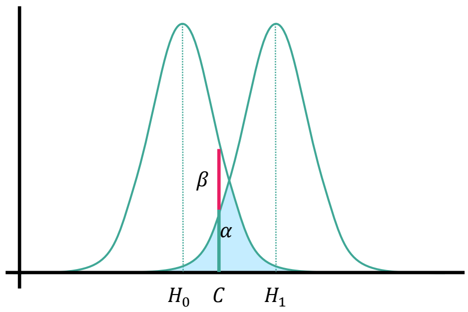
במבחני השערות מדגמים גדולים (מעל 30 דגימות ומעלה) מספר המדדים המתפלגים נורמלית במדגם גדול. ולכן, אפשר להשתמש בחישובים שנעשו על מנת לבדוק מובהקות סטטיסטית. אם קטן, יש לבצע תיקונים.

בהינתן n דגימות בלתי תלויות של משתנה נורמלי נחשב- $t = \frac{E(X)-\mu}{\sigma/n}$. כאשר, הנתונים מתפלגים עם מספר דרגות חופש השווה ל- $n-1$. עבור n גדול ההתפלגות של t שואפת להתפלגות נורמלית.

מציאת B –

B היא ההסתברות לטעות מסוג 2

1. מציאת הערך Z_a
2. מתקננים את C על מנת למצוא את β. $\beta = 1 - \underbrace{P(Z < \frac{c-\mu_{H_1}}{S_{H_1}/\sqrt{n}})}_{\text{לנקודה משמאל הנקודה } C}$



מבחני השערות:

מבחני השערות במדגמים גדולים: ממוצעים

נגדיר -

- $E[X]$ - ממוצע המדגם
- μ - התוחלת של האוכלוסייה
- σ - סטיית התקן של האוכלוסייה
- n – גודל המדגם.

המשתנה המתוקנן יהיה: $Z = \frac{E(X)-\mu}{\sigma/\sqrt{n}}$ ולא נדחה את השערת האפס כאשר $-\Phi(a) < Z < \Phi(a)$ ונדחה אם לא מתקיים.

מבחני השערות במדגמים גדולים: הצלחות

נגדיר -

- $\mu_P = p = \frac{\text{הצלחות}}{n}$ - תוחלת ההצלחות באוכלוסייה
- $\sigma_P = \sqrt{p(1-p)/n}$,

הערך הקריטי יהיה $C = P + Z_a \cdot \sqrt{p(1-p)/n}$

המשתנה המתוקנן יהיה: $Z = \frac{P-p}{\sqrt{p(1-p)/n}}$ ולא נדחה את השערת האפס כאשר $-\Phi(a) < Z < \Phi(a)$ ונדחה אם לא מתקיים.

מבחני השערות במדגמים גדולים: הפרש בין ממוצעים

נגדיר –

- $E[X_1], E[X_2]$ – ממוצעי המדגמים הבלתי תלויים בגדלים n_1 ו- n_2 בהתאמה.
- μ_1, μ_2 – ממוצעי האוכלוסיות בהתאמה.
- σ_1, σ_2 – סטיות התקן של האוכלוסיות בהתאמה.

נגדיר הנחות-

- **H_0** - אין הבדל בין הממוצעים $\mu_1 = \mu_2$ כלומר: $H_0: \mu = \mu_1 - \mu_2 = 0$
- **H_1** - קיים הבדל בין הממוצעים $\mu_1 \neq \mu_2$ כלומר: $H_1: \mu = \mu_1 - \mu_2 \neq 0$

נוכל להשתמש בערך הקריטי- $C = \mu_1 - \mu_2 + Z_a \cdot \sigma_{X1-X_2}$ ומשווים מול $\overline{X_1} - \overline{X_2}$.

סטיית התקן בין הפרשי המדגם - $E^2_{S1-S2} = E((X_1 - X_2)^2) = E(X_1^2) - E(X_2^2)$

שגיאת התקן של ההפרש בין המדגמים - $\sigma_{S1-S2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

המשתנה מתוקנן הוא - $Z = \frac{(E(X_1)-E(X_2))-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ ותחת הנחת האפס - $Z = \frac{E(X_1)-E(X_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$.

מבחני השערות במדגמים גדולים: הפרש בין הצלחות

נרצה להמיר את ההפרש בין הממוצעים לבחינה אם מדגם של הפרש שונה מאפס. נגדיר: $\hat{p} = \frac{p_1 \cdot n_1 + p_2 \cdot n_2}{n_1 + n_2}$

נקבל שסטיית התקן היא - $\sigma_{P_1-P_2} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

הערך הקריטי יהיה $C = P_1 - P_2 + Z_a \cdot \sigma_{P_1-P_2}$

ולכן השונות המתוקננת היא - $Z = \frac{P_1-P_2}{\sigma_{P_1-P_2}}$

מבחני השערות במדגמים גדולים: מבחנים של זוגות

בהנחה שיש זוגות של שתי מדידות שונות $(x_1, y_1), \ \ldots, \ (x_n, \ y_n)$ נחשב את ההפרשים $D_1 \ldots D_n$.
לסדרת ההפרשים ניתן לחשב תוחלת וסטיית תקן משלהם. ככה נוכל לטעון:

- **H_0** - אין הבדל בין הממוצעים $\mu_X = \mu_Y$ כלומר: $H_0: \mu = \mu_X - \mu_Y = 0$.
- **H_1** - קיים הבדל בין הממוצעים $\mu_X \neq \mu_Y$ כלומר: $H_1: \mu = \mu_X - \mu_Y \neq 0$.

נגדיר את סטיית התקן - $S_D = \sqrt{\frac{\sum_{i=1}^{n}((X_i - Y_i) - \bar{X}_D)^2}{n-1}}$

נוכל להשתמש בו במבחן הקריטי באופן הבא: $C = \mu_D + t_a \cdot \frac{S_D}{\sqrt{n}}$

ונוכל לתקנן כך ש- $t = \frac{X_D - \mu_D}{S_D/\sqrt{n}}$

הערות:

1. כאשר השונות לא ידועה ו- $n \leq 30$ נעזר ב- t עם $n-1$ דרגות חופש, כאשר משווים בין שני משתנים אז דרגות החופש תהיה $n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$.
2. בתקנון הראתי פה רק עבור מבחן ימני, בעבור מבחן שמאלי ודו צדדי לשנות בהתאם.

תעבדו עם מה שנוח לכם, השתדלתי לתת ייצוג לכל השיטות.

מבחנים א-פרמטריים (Non-parametric)

מבחנים שלא מניחים הנחות על התפלגות הנתונים – לכן הם פחות רגישים לנתונים יוצאי דופן, לכן, אין צורך לבדוק את ההתפלגות של הנתונים, עובדים על נתונים אורדינליים וקטגורים.

במצב כזה נבחן את היחס בין ההסתברות הנצפית להסתברות התיאורטית. אפשר לרשום את ההסתברות למדגם שקיבלנו, תחת ההתפלגות התיאורטית. נגדיר מדד χ^2 באופן הבא:

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

- מתפלג בהתפלגות χ^2 עם k-1 דרגות חופש (k הוא מספר הספרות)
- **E_i** - התוצאה הצפויה.
- **O_i –** התוצאה שקיימת.

נשים לב שככל ש- χ^2 גדול יותר, כך יש אי התאמה גדולה יותר. לכן, מספיק שספרה אחת תהיה מאד שונה כדי שכל המבחן יהיה מאוד שונה (כמו שרצינו).

מבחן טיב ההתאמה

בדיקת התאמה **עבור משתנה בודד** בין ההתפלגות הקיימת להתפלגות הצפויה, למשל בין ניסוי והתוצאות של הניסוי. המבחן יחושב כך:

1. הגדרת טבלה של ערך קיים וערך מצופה, הערך המצופה יילקח מהאחוז או ממה שצופה.
2. חישוב כל ההפרשים
3. מציאת ערך $\chi^2_{statistic}$ באמצעות ההפרשים.
4. מציאת הערך $\chi^2_{Critical} = \chi^2(df = n-1, \alpha)$ מתוך טבלת ה- χ^2.
5. על פי המבחן החד צדדי לראות האם מתקיים האי שוויון ולקבל את הטענה.

מבחן אי תלות –

במידה ונרצה לראות האם יש תלות בין **שני** משתנים קטגורים, כמו עיר מגורים והשכלה.

המבחן יחושב כך:

1. הגדרת טבלה של ערכים קיימים עם סכומי עמודה, שורה וסכום כולל.
2. הגדרת טבלה של ערכים מצופים כך: $E_{i,j} = \frac{rowTotal \cdot colTotal}{total}$
3. מציאת ערך $\chi^2_{statistic} = \sum_{i=0}^{rows} \sum_{j=0}^{cols} \frac{\left(E_{i,j} - O_{i,j}\right)^2}{E_{i,j}}$
4. מציאת הערך $\chi^2_{Critical} = \chi^2(df = (r-1)\cdot(c-1), \alpha)$ מתוך טבלת ה- χ^2.
5. על פי המבחן החד צדדי לראות האם מתקיים האי שוויון ולקבל את הטענה.

מדידת הפרשים א-פרמטרית במדגמים בלתי תלויים Mann-Whitney U Test

עבור דגימות X, Y **התפלגויות** אורדינליות רציפות - F_X, F_Y **בלתי תלויות**, והתפלגות **אוכלוסייה לא ידועה**, שימושי עבור השוואה של שביעות רצון בין שתי קבוצות, מאחר שרציפות $P(X = Y) = 0$ אז:

- **H_0** - $P(x > y) = P(x < y) = \frac{1}{2}$
- **H_1** - $P(x > y) \neq P(x < y)$

נחשב את ה- U statistic: זהו מדד סטטיסטי שמחשב את ההפרדה או ההבדל בין שתי קבוצות

מדגם ללא הנחת נורמליות

$$U_i = \underbrace{\sum_{j=1}^{n_i} R_{i,j}}_{\text{מיקום לפי הסדר}} - \overbrace{j}^{\text{המיקום המוקדם ביותר}} = R_i - \underbrace{\frac{n_i(n_i+1)}{2}}_{\sum_{j=1}^{n} j = \frac{n(n+1)}{2}}$$

המבחן יחושב כך:
1. דירוג כלל הנתונים בקבוצה מהקטן לגדול.
2. דירוג כלל האיברים, אם יש איברים בעלי ערך זהה אז נבחר את הממוצע של הדירוג המתאים.
   כדאי מאוד לבדוק בסוף את הדירוג כך שסכום הדירוגים שווה ל- $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$.
3. נחשב את סכום הדרגות לכל אחת מהקבוצות $U_x = \sum_{j=1}^{n_x} R_{x,j}$,$U_y = \sum_{j=1}^{n_y} R_{y,j}$.
4. נגדיר $U = \min\left(U_x, U_y\right)$.
5. בטבלת U נמצא את הערך הקריטי עבור מספר העמודות ומספר השורות.
6. בחירת ערך קריטי עבור $U_{critical} = U\left(n_x, n_y\right)$ מטבלת Mann-Whitney U
7. אם - $U < U_{critical}$ נדחה את H_0.

Kruskal-

עבור התפלגות **סטטיסטית לא ידועה**, עבור **יותר מ-** $2 < k$ **קבוצות בלתי תלויות**. נרצה לבדוק איזה מהקבוצות הן נבדלות . לרמת מובהקות $a_{adjusted} = \frac{a_{original}}{num\ of\ comparisons}$ צמדים יש לנו וזהו יהיה מספר ההשוואות $\binom{k}{2}$, ועל כל קבוצה כזו נבצע Mann-Whitney U Test רגיל. (זו בעצם הרחבה של Mann-Whitney U Test על ה- $a_{adjusted}$ שמצאנו).

מדידת הפרשים א-פרמטרית במדגם מזווג: Wilcoxon sign-rank test

עבור **שני מדגמים מזווגים** תואמים, **<u>תלויים</u>** כאשר x_i, y_i **לא מתקיימת הנחת התפלגות אוכלוסייה ידועה** ונרצה לבדוק אם החציון של x_i שונה מהחציון של y_i נסמן את החציון של x,y ב- m_x, m_y בהתאמה. משמש עבור בדיקה של שתי קבוצות תואמות, ההשערות הן-

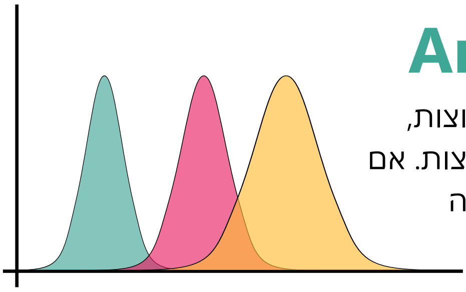
- ***H_0*** - $m_x = m_y$
- ***H_1*** - $m_x \neq m_y$

נניח שהדגימות x, y הן מהתפלגויות אורדינליות כאשר המבחן מבוסס על ההפרשים בין הנקודות.

המבחן יחושב כך:
1. חישוב כלל ההפרשים בין שתי הקבוצות.
2. סידור הערך המוחלט של ההפרשים מהנמוך לגבוהה בדומה ל- Mann-Whitney U Test.
3. הוספת הסימנים + או – בהתאם.
4. חישוב סכום המיקומים של כל ההפרשים החיוביים והשליליים W^+, W^- ונבחר את המינימלי מבניהם $W = \min\left(W^+, W^-\right)$.
5. נגיע לטבלת Wilcoxon ונחפש את הערך הקריטי בו $df = n$, עם ערך המדגם קטן מהערך הקריטי נדחה את H_0.

היחס בין χ^2, Z, t-

מתקיים ש- $t(df, a) = \frac{N(0,1)}{\sqrt{\chi^2(df,a)/df}}$. לא ראינו את זה באף מקום מלבד שיעורי הבית, כדאי לזכור ולשים בדף נוסחאות.



Analysis of Variance - ANOVA

הרעיון של המבחן הוא להשוות בין פיזור הנקודות עבור יותר מ- 2 קבוצות, המתפלגות נורמלית בתוך כל הקבוצות לבין פיזור הנקודות בין הקבוצות. אם הפיזור בין הקבוצות גדול יותר מבין זה שבתוכן אז הממוצע שלהן שונה באופן משמעותי.

הנחות יסוד הן שההתפלגות של כל קבוצה היא נורמלית ושכל הדגימות הן בלתי תלויות ושוות התפלגות. והשונות היא ההבדל בין הקבוצות.

במבחן, נשווה בין פיזור הנקודות בתוך כל קבוצה לבין הפיזור בין שאר הקבוצה.

מבחן ANOVA למשתנה אחד (One way ANOVA)

נסמן ב-μ_i את התוחלת של המשתנה ה-i. ונתונות n דגימות מ-q קבוצות $(n = n_1 + n_2 + \cdots + n_q)$

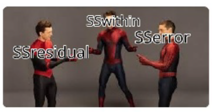
- H_0 - $\mu_j = \mu_k \quad \forall j, k$
- H_1 - $\exists j, k:\ \mu_j \neq \mu_k$

כמו כן נרשום $y_{i,j} = \mu_j + \epsilon_{i,j}$ כאשר $\epsilon_{i,j} \sim N(0, \sigma)$ (סטית התקן אינה ידועה).

- ממוצע הנתונים בתוך קבוצה j ∈ {1,2 … q} $\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{i,j}$.
- הממוצע של כלל הנקודות יהיה - $\bar{y} = \frac{1}{n}\sum_{j=1}^{q}\sum_{i=1}^{n_q} y_{i,j} = \frac{1}{n}\sum_{j=1}^{q} n_j \bar{y}_j$.
- פיזור הנקודות סביב הממוצע הכללי – $SS_{Total} = \sum_{j=1}^{q}\sum_{i=1}^{n_j}\left(y_{i,j} - \bar{y}\right)^2$
- ונקבל - $\sum_{i=1}^{n_j}\left(y_{i,j} - \bar{y}\right)^2 = \sum_{i=1}^{n_j}\left(y_{i,j} - \bar{y}_j\right)^2 + n_j\left(\bar{y}_j - \bar{y}\right)^2 = \boldsymbol{(n_j - 1)s_j^2 + n_j\left(\bar{y}_j - \bar{y}\right)^2}$

נקבל את הכולל - $SS_{Total} = \sum_{j=1}^{q}\left[(n_j - 1)s_j^2 + n_j\left(\bar{y}_j - \bar{\bar{y}}\right)^2\right]$ נפרק את המבחן:

- $\boldsymbol{SS_{Total} = \sum_{j=1}^{q}\left[(n_j - 1)s_j^2 + n_j\left(\bar{y}_j - \bar{\bar{y}}\right)^2\right]}$– סכום הריבועים הכולל של כל האיברים .
- $\boldsymbol{SS_{Within} = \sum_{j=1}^{q}\sum_{i=1}^{n_q}\left(y_{i,j} - \bar{y}_j\right)^2 = \sum_{j=1}^{q}(n_j - 1)s_j^2}$– סכום השתנות בתוך הקבוצות.
- $\boldsymbol{SS_{Between} = \sum_{j=1}^{q} n_j\left(\bar{y}_j - \bar{\bar{y}}\right)^2}$– ההשתנות בין הממוצעים של הקבוצות השונות.



סה"כ: $SS_{Total} = SS_{Within} + SS_{Between}$ כלומר הפיזור נקודות סביב הממוצע הכללי נובע מהפיזור סביב הממוצע של כלל הקבוצה והממוצע של כל קבוצה סביב הממוצע הכללי.

מקור השונות	ממוצע סכום הריבועים	סכום הריבועים	דרגות חופש
בין הקבוצות	$s_{Between}^2 = SS_{Between}/(q-1)$	$SS_{Between}$	$df_{Between} = q - 1$
בתוך הקבוצות	$s_{Within}^2 = SS_{Within}/(n-q)$	SS_{Within}	$df_{Within} = n - q$
סה"כ	$\sum_{j=1}^{q}\left[(n_j - 1)s_j^2 + n_j\left(\bar{y}_j - \bar{y}\right)^2\right]$	SS_{Total}	$df_{Total} = n - 1$

נבצע בחינה בין היחס בין פיזור הנקודות בתוך הקבוצות לבין פיזור הנקודות בין הקבוצות, סה"כ:

$$MS_B = \frac{SS_{Between}}{df_{Between}}$$

$$MS_W = \frac{SS_{Within}}{df_{Within}}$$

$$F = \frac{s_{Between}^2}{s_{Within}^2} = \frac{SS_{Between}/(q-1)}{SS_{Within}/(n-q)} = \frac{MS_B}{MS_W}$$

שלבים לביצוע מבחן One Way ANOVA –

1. חישוב ממוצע לכל קבוצה.
2. נחשב ממוצע של כל הקבוצות ביחד.
3. נמצא את סכום השונויות בתוך כל הקבוצות - $SS_{Within}=\sum_{j=1}^{q}\sum_{i=1}^{n_q}\left(y_{i,j}-\bar{y}_j\right)^2$.
4. נמצא את סכום השונויות בין הקבוצות - $SS_{Between}=\sum_{j=1}^{q}n_j(\bar{y}_j-\bar{y})^2$.
5. נמצא את דרגות החופש - $df_{Between}=q-1,\ df_{Within}=n-q, df_{Total}=n-1$.
6. למצוא את הממוצע שונות בין הקבוצות MS_B עבור $SS_{Between}$. $MS_B=\frac{SS_{Between}}{df_{Between}}$.
7. חישוב ממוצע השונות בתוך הקבוצות MS_W עבור SS_{Within}. $MS_W=\frac{SS_{Within}}{df_{Within}}$.
8. נחשב את ה- $F_{Statistic}$ סטטיסטי מדגם - $F_{Statistic}=\frac{MS_B}{MS_W}$.
9. חישוב F_{Critic} מהטבלת F באופן הבא- $F_a\left(\overbrace{q-1}^{df_b},\underbrace{n-q}_{df_w}\right)$.
10. אם $F_{Statistic}$- קטן מ- F_{Critic} נדחה את H_0.

ANOVA בשני משתנים

שלבים לביצוע מבחן Two Way ANOVA –

בהינתן שני משתנים בת"ל- עבור משתנה A בגודל p ומשתנה B בגודל q.

1. נגדיר מערכת השוואות לשני משתנים בנפרד ומערכת אחת על השפעה בין שני הגורמים.
2. נחשב ממוצעים עבור כל משתנה ועבור אינטראקציית משתנים. מומלץ לארגן בטבלה.

חישוב כולל -

3. נחשב ממוצע לכל הערכים $\bar{y}_{total}$.
4. נחשב את SS_{Total} כאשר $SS_{Total}=\sum_{m=1}^{p}\sum_{i=1}^{q}\sum_{j=1}^{n_{i,j}}\left(x_{m,i,j}-\bar{y}_{total}\right)^2$.
5. חישוב דרגות החופש עבור A, B והכולל - $df_{Total}=n\cdot p\cdot q-1\ ,\ df_B=q-1\ ,\ df_A=p-1$.
6. חישוב $MS_T=\frac{SS_{Total}}{df_{Total}}$

חישוב עבור A ו- B -

7. חישוב ה- SS_A וה- SS_B - $SS_A=\sum_{i=1}^{p}n_i(\bar{A}_i-\bar{y}_{total})^2$, $SS_B=\sum_{i=1}^{p}n_i(\bar{A}_i-\bar{y}_{total})^2$
8. חישוב MS_A וה- MS_B - $MS_A=\frac{SS_A}{df_A}$ ו- $MS_B=\frac{SS_B}{df_B}$

עבור בתוך D- Between AB

9. חישוב ה- SS_{AB} - $SS_{AB}=SS_{Between}=\sum_{i=1}^{p}\sum_{j=1}^{q}\left(\overline{A_iB_j}-\bar{A}_i-\bar{B}_i+\bar{y}_{total}\right)^2$
10. חישוב df_{AB} - $df_{AB}=(p-1)(q-1)$
11. חישוב MS_{AB} - $MS_{AB}=\frac{SS_{AB}}{df_{AB}}$

עבור השגיאה Within AB.

12. חישוב ה- SS_{Within} - $SS_{Within}=\sum_{m=1}^{p}\sum_{i=1}^{q}\sum_{j=1}^{n_{i,j}}\left(x_{m,i,j}-\overline{AB}_{i,j}\right)^2$
13. חישוב df_{Within} - $df_{Within}=(n-1)\cdot p\cdot q$
14. חישוב MS_{Within} - $MS_{Within}=\frac{SS_{Within}}{df_{Within}}$

חישוב F וקבלה או שלילה -

15. נחשב את F_A, F_B ו- F_{AB} - $F_A=\frac{MS_A}{MS_{Within}}, F_B=\frac{MS_B}{MS_{Within}}, F_{AB}=\frac{MS_{AB}}{MS_{Within}}$
16. נחשב F_{critic} עבור F_A, F_B ו- F_{AB} באופן זהה: - $F_x(df_x, df_{Within})$
17. אם $F_{Statistic}$- קטן מ- F_{Critic} נדחה את H_0.
18. במידה ודחינו את H_0, נרצה לדעת איזה קבוצה שונה מהאחרות, נשתמש במבחן post hoc.

כאשר יש יותר מגורם בלתי תלוי אחד

למשל השוואה בין קבוצות לפי שני קריטריונים, אז נגדיר:

$x_{j,k} =$	$\mu +$	$\alpha_j +$	$\beta_k +$	$\Delta_{j,k}$
הדרך להביע את ההבדל	ממוצע האוכלוסיה	שינוי כתוצאה מטיפול 1	שינוי כתוצאה מטיפול 2	שונות סביב הממוצעים

Kruskal-Wallis

כאשר הנתונים לא מתפלגים נורמלית נפנה למבחן - Kruskal-Wallis (מבחן Rank-Sum). נחליף את הערכים בסדר שלהם במדגם.

נשים לב ש- SS_{Within} לא במכנה כיוון שלא נרצה להניח כלום על ההתפלגות בתוך הקבוצות. כמו כן, אם יש לפחות 5 דגימות בכל קבוצה אז ההתפלגות היא בקירוב - χ^2_{q-1} ואם אין אף ערך שחוזר על עצמו אז $SSR_{Total} = \frac{(n-1)n(n+1)}{12}$.

המבחן יחושב כך:

1. דירוג הנתונים מהקטן לגדול והחלפת כל איבר בדירוגו כמו ב- Mann-Whitney U Test.
2. נגדיר סטטיסטי - $\chi^2_{Statistic} = H = \frac{SSR_{Between}}{SSR_{Total}/(n-1)} = \frac{\sum_{i=1}^{q} n_i(\bar{r}_i-\bar{\bar{r}})^2}{\sum_{i=1}^{q}\sum_{j=1}^{n_i}\left(r_{i,j}-\bar{\bar{r}}\right)^2/(n-1)}$ עבור q קבוצות.
3. בטבלה של χ^2 נמצא את הערך הקריטי עם $\chi^2_{Critic}(df = k-1)$.
4. השוואת ה- χ^2_{Critic} ל- $\chi^2_{Statistic}$ לפי המבחן הצדדי וקבלה או דחייה של H_0.

Tukey's Honestly Significant Difference (HSD) Test(Post-hoc tests)

נרצה לדעת איזו שונות וממי היא מבין הקבוצות. במידה וה- ANOVA אמרה שיש הבדל משמעותי בין שתי הקבוצות אנחנו לא יודעים אילו קבוצות שונות, רק שיש זוג אחד ששונה. להשוות בין כל הקבוצות עלול לקחת המון השוואות, דבר שיעלה את הסיכוי לשגיאה מסוג ראשון. לכן, נרצה להבין אילו מהקבוצות שונות באופן משמעותי, נשתמש ב- Post hoc test הדואג למנוע שגיאות כאלה. משווה כל זוג של ממוצעים, כלומר, $\mu_i - \mu_j$ לכל i, j. נניח שהנתונים בלתי תלויים, הנתונים מתפלגים נורמלית ושהשונות בקבוצות דומה. המבחן: $q_s = \frac{|\mu_A - \mu_B|}{SE}$.

פילוג q –

Studentized range distribution, נדגום n דגימות מ-k אוכלוסיות עם התפלגות זהה $N(\mu, \sigma)$. ונגדיר:

- $\bar{x}_{min}$ את הממוצע של הקטן ביותר (של אחת מהקבוצות)
- $\bar{x}_{max}$ את הממוצע הגדול ביותר (של אחת מהקבוצות)
- s^2 את סטיית התקן של כלל הנקודות.

אזי $q = \frac{\bar{x}_{max} - \bar{x}_{min}}{s/\sqrt{n}}$ מתפלג בהתפלגות q.

הפרש הממוצעים המינימלי שאינו מובהק סטטיסטית –

אם $q_s = \frac{|\mu_A - \mu_B|}{SE}$. נשתמש בפילוג q המכונה Studentized range distribution. כיוון ש- $\frac{|\mu_A - \mu_B|}{SE}$ נקבל ש- $\text{HSD} = |\mu_A - \mu_B| = q \cdot SE = q\sqrt{\frac{SE^2}{n}}$

המבחן יחושב כך:

עבור n מספר האיברים בכל קבוצה, k מספר הקבוצות ו- q – ההתפלגות (Z או t בהתאם לגודל המדגם)

1. חישוב ANOVA על מנת למצוא אם קיבלנו את H_1
2. חישוב הערך הקריטי של q $HSD = q_a\sqrt{\frac{2 \cdot MSE}{n}}$ עבור 30 ומעלה נשתמש בטבלת Z.
3. נחשב את הבדלי הממוצעים בין כל זוג קבוצות. אם ההפרש בין שני הממוצעים גדול מה- HSD אז יש ביניהם הבדל מובהק.

רגרסיה –

סימונים –

- מדגם ובו זוגות של נקודות x_i, y_i שנרצה לקרב ע"י ישר.
- **X –** יהיו משתנים בלתי תלויים.
- **Y –** יהיו משתנים תלויים.
- משוואת הישר היא - $y = ax + b + \varepsilon$
- הערך על הישר בנקודה x_0 יהיה $(x_0, \widehat{y_0})$
- המרחק הריבועי של נקודה על הישר מהנקודה המתאימה במדגם הוא $\epsilon_i = (y_i - \hat{y}_i)^2$ כלומר $\hat{y}_i$ זה החיזוי של הערך על הישר בנקודה x_i לפי הרגרסיה.

מציאת קו ישר -

$$E = \min(\epsilon) = \min_{w,b}(\sum_i (y_i - w \cdot x_i - b)^2)$$

$$E = \min(\epsilon) = \min(\sum_i \epsilon_i) = \min(\sum_i (y_i - \hat{y}_i)^2)$$

לאחר פישוטים נקבל שהמינימום הוא- אפרנטלי זה נקרא שיטת הריבועים המינימליים שתכירו את המונח.

$$b = \beta_0 = \frac{\bar{y} \cdot \Sigma x_i^2 - \bar{x} \cdot \Sigma y_i \cdot x_i}{\Sigma x_i^2 - \frac{1}{n} \cdot (\Sigma x_i)^2}$$

$$a = \beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\Sigma x_i y_i - n \cdot \bar{x} \cdot \bar{y}}{\Sigma x_i^2 - n\bar{x}^2}$$

ניתן למצוא את a להציב בנוסחת הקו הישר $y = ax + b$ את $x = \bar{x}, y = \bar{y}$. אם שואלים אותנו מה חזוי, פשוט מציבים את ה- x ומקבלים את ה- $\hat{y}$ החזוי.

באמצעות אלגברה ליניארית-

נגדיר n דוגמאות (תצפיות) ו - m משתנים (תכונות)

- **X -** היא מטריצה בגודל $n \times (m+1)$ הנראית כך:
- **W -** הוא ווקטור התוצאה בגודל $(m+1) \times 1$
- **Y -** היא ווקטור משקלים $n \times 1$ הנראית כך:

$$X \cdot W = Y \to X = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{m,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{m,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{m,n} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad W = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_m \end{bmatrix}$$

- ועבור כל $0 \le i \le n$ יתקיים: $y_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_m x_{i,m}$ את התוחלת נקבל כך- $\bar{y} = \frac{1}{n}\Sigma_i^n w_0 + w_1 x_i + \epsilon_i$

נגזרת של מטריצות

אבל אנחנו רוצים למצוא את W, נעשה גזירה של מטריצות:

$$\frac{\partial y}{\partial x} = A = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

- ונניח ש- $y = Ax$
- כאשר y בגודל $m \times 1$
- x , בגודל $n \times 1$,
- A בגודל $n \times m$.

צ"ל:

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

הוכחה: אם A מטריצה סימטרית מתקיים $\frac{\partial}{\partial z}(z^T Az) = 2Az$ מאחר ש- $z^T Az = \sum_{i=1}^n \sum_{j=1}^n z_i A_{ij} z_j$,, הנגזרת לפי z: $\frac{\partial(z^T Az)}{\partial z_k} = \sum_{i=1}^n z_i A_{ik} + \sum_{j=1}^n A_{kj} z_j$. אז אם נרצה למזער את $E = (Y - XW)^T(Y - XW)$ נפתח ונקבל ש- $E = Y^T Y - Y^T XW - W^T X^T Y + W^T X^T XW = Y^T Y - 2Y^T XW + W^T X^T XW$ נגזור ונשווה לאפס ונקבל- $\frac{dE}{dW} = 2Y^T X - 2W^T X^T X = 0 \Rightarrow X^T XW = X^T Y \Rightarrow$ אז: $(X^T X)^{-1} X^T XW = (X^T X)^{-1} X^T Y$ סך הכל יצא- $\boldsymbol{W = \left(X^T X\right)^{-1} X^T Y}$ אפשר ליישם עם המחשבון יש מדריך בסוף.

מקדם המתאם - מתאם פירסון Pearson -

$$R = r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n\cdot\sigma_x\cdot\sigma_y} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

$-1 \le r \le 1$ והוא סימטרי עבור x ו- y. $R = r*$

- $SS_{explaned}$ – החלק של השונות שתלוי ב- X. $SS_{Reg} = S_{explaned} = \sum(\hat{y}_i-\bar{y})^2$
- $SS_{residual}$ – החלק בשונות שלא תלוי ב-X, המינימום שגיאות $SS_{residual} = SS_{error} = \sum(y_i-\hat{y}_i)^2$.
- SS_{Total} – סכום הריבועים של השונות של הנתונים המקוריים.$S_{Total} = \sum(y_i-\bar{y})^2$

נקבל גם ש- $$\frac{SS_{explaned}}{SS_{Total}} = \frac{\overbrace{E(\hat{y}-\bar{y})^2}^{\hat{y}=wx+b}}{E(y-\bar{y})^2} = \frac{\overbrace{E((b+wx-\bar{y})^2)}^{\bar{y}=w\bar{x}+b}}{\sigma_y^2} = \frac{\overbrace{w^2E(x-\bar{x})^2}^{var(ax)=a^2\cdot var(x)}}{\sigma_y^2} = \frac{\overbrace{w^2}^{w=\frac{\sigma_{xy}}{\sigma_x^2}}\ \sigma_x^2}{\sigma_y^2} = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} = r^2 = R^2$$

לכן, נתייחס ל- R^2 בתור אחוז השבר של השונות המוסברת ע"י הרגרסיה. נשים לב גם שמתקיים - $S_{Total} = S_{explaned} + S_{residual}$ לכן ניתן לכתוב גם - $R^2 = \frac{SS_{Total}-S_{residual}}{SS_{Total}}$

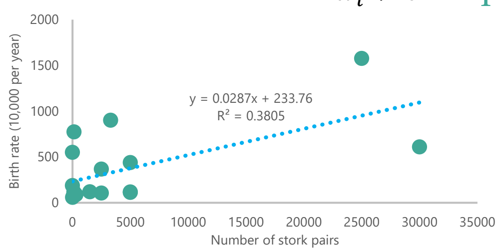
שאריות ברגרסיה

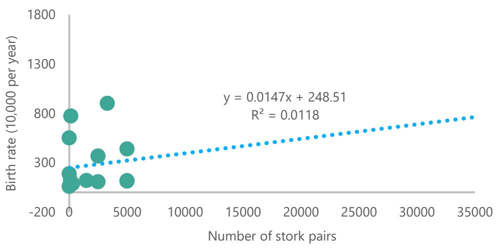
הבדל חשוב הוא ש- r מייצג את הקשר בין שני המשתנים והכיוון בין שני משתנים, R^2 מייצג את כמות השונות של y שהוסברה על ידי x כלומר אם יש לנו $R^2 = 0.17$ אז X מסביר 17% מ-Y , ו- a, w_i, β_i הם שיפוע והמשתנה החופשי של הקו רגרסיה המייצג השפעה כמותית העוזרת בחיזוי.

חישוב מובהקות סטטיסטית –

האם השיפוע של המקדם משמש מובהקות סטטיסטית? נגדיר:

- H_0 - $w_i = 0$
- H_1 - $w_i \ne 0$





- השונות של השארית - $\sigma^2 = \frac{\|y-XW\|^2}{n-k} = \frac{RSS}{n-k}$
- ברשום מטריצה - $SE(W) = \sqrt{diag(\sigma^2(X^TX)^{-1})}$
- לכן - $t = \frac{\hat{w}}{SE(\hat{w})}$ כשיש $(n-k-1)$ דרגות חופש.
- **רגרסיה רב מימדית –** $Y = X\cdot W$
- **רגרסיה לא לינארית –** $y = w_1\cdot x + w_2\cdot w^2$

מציאת מובהקות עבור שיפוע β_i בקו הרגרסיה – עבור חיזוי.

1. H_0 – אין מובהקות סטטיסטית כלומר $\beta_i = 0$, H_1 קיים קשר $\beta_i \ne 0$.
2. חישוב הסטטיסטי $t = \frac{\beta_i}{S_b}$ כש: $S_b = \sqrt{\frac{\sum_{i=1}^n(y_i-\hat{y}_i)^2}{(n-2)\sum_{i=1}^n(x_i-\bar{x}_i)^2}}$
3. חישוב הנקודה הקריטית t_{critic} כך: $t_{critic} = \left(\frac{a}{2}, df = n-k-1\right)$.
4. אם מתקיים $-t_{critic} < t < t_{critic}$ נקבל את טענת האפס, אחרת נדחה.

מציאת מובהקות עבור מתאם המדגם r – עבור היחס.

1. נגדיר H_0 אין קשר בין המשתנים כלומר $r=0$, H_1 קיים קשר $r \neq 0$.
2. נחשב את הסטטיסטי $\theta = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$
3. חישוב הנקודה הקריטית t_{critic} כך: $t_{critic} = \left(\frac{a}{2}, df = n-2\right)$.
4. אם מתקיים $-t_{critic} < t < t_{critic}$ נקבל את טענת האפס, אחרת נדחה.

רגרסיה של סדר (Rank) –

1. נדרג את ערכי X וערכי Y בסדר עולה בנפרד בדומה ל- Mann-Whitney.
2. לכל נקודה (x_i, y_i) נחשב את ההפרש $d_i = |x_i - y_i|$
3. נחשב $r_{rank} = 1 - \frac{6\sum d_i^2}{n(n-1)}$

תיקון R^2 לגודל המדגם: $Adjusted\ R^2$

ככל שנוסיף יותר משתנים למודל הוא ייטה להתאים יותר לנתונים ולכן R^2 צפוי לגדול. נגדיר:

- n - מספר הנקודות במדגם
- p - מספר המשתנים הבלתי תלויים.

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

- $\bar{R}^2$ יכול להיות קטן מ- 0.
- תמיד מתקיים ש- $\bar{R}^2 \leq R^2$.
- $\bar{R}^2$ עולה בהוספת משתנים רק אם העלייה ב-R^2 גדולה מזו הצפויה באופן אקראי.

Identifiability –

אם $X \in \mathbb{R}^{n \times p}$ ראינו ש- $rank(X) = p$ במידה ו- $p > n$ נאמר שלא ניתן לזהות את w (not identifiable). אפשר לבצע חיזוי אבל המודל חסר ערך. זה עלול לקרות כאשר יש קורלציה בין משתנים, כאשר משתנים זהים מופיעים במודל. יש פחות נתונים ממשתנים. במצבים כאלה לא נוכל להחליט איזה מהמשתנים נמצא בקורלציה עם המשתנה התלוי.

VIF-Variance Inflation Factor

לכן, אם מנסים להעריך כמה המצב חמור ניתן להשתמש ב- Variance Inflation Factor זוהי דרך לחשב עד כמה השונות של מקדם רגרסיה גדול מערכו האמיתי כתוצאה מקולורציה. בעת חישוב של מודל רגרסיה - $y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n$, נבצע את השלבים הבאים:

1. נבנה n מודלי רגרסיה כשבכל פעם אחד המשתנים הוא תלוי: $x_1 = \alpha_0 + \alpha_2 \cdot x_2 + \cdots + \alpha_n \cdot x_n$
2. נסמן ב- R_i^2 את ה- R של המודל ה-i
3. נחשב: $VIF_i = \frac{1}{1-R_i^2}$
4. נבדוק ערכים:
   - אם $VIF = 1$ אין בעיה
   - אם גדול מ- 5 אז יש בעיה שנקראת collinearity
   - אם גדול מ- 10 אז יש בעיה משמעותית.

כאשר יש בעיה נוציא את המשתנים הבלתי תלויים מהמודל או להשתמש בשיטות כמו PCA על מנת למצוא הטלה שונה. או שימוש בשיטת רגרסיה שכופות שימוש בפחות משתנים.

Ridge regression
במידה ו- p > $rank(X)$ אז אפשר להכניס אילוץ לפונקציה על מנת שהמטריצה X^TX תהיה הפיכה. נעשה זאת על ידי מזעור של השגיאה הריבועית תחת האילוץ ש- $W^TW = c$ פרמטר לבחירתנו, נשתמש בכפלי לאגרנג' ונרשום את הפונקציה כך: $\lambda \geq 0$ כופל לגראנג'.

$$\min_{\beta}(y - X \cdot W)^T(y - X \cdot W) + \lambda(W^TW - c).$$

Lasso
דרך שונה לאילוץ מודלים, היא לתת הקצאה לפרמטרים, כלומר לבחור רק את הערכים שנמצאים מתחת לערך t כלשהו: $\min_{W} \sum(y_i - w_i \cdot x_i)^2 \ subject\ to\ \sum|w_i| \leq t$

דרך דומה: **Elastic nets**:

$$\min_{W}\{\|y - X \cdot W\|_2^2 + \lambda_1\|W\|_1 + \lambda_2\|W\|_2^2\}$$

$$\min_{W} \sum(y_i - w_i \cdot x_i)^2 \quad subject\ to\ (1-\alpha)\|W\|_1 + \alpha\|W\|_2^2 \leq t$$

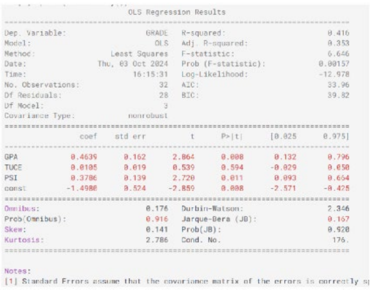
Stepwise regression
שיטה לבניית מודל רגרסיה שאינו משתמש בכל המשתנים:

- Forward selection
1. מתחיל במודל ריק
2. מוסיפים את הערכים המשמעותיים ביותר (בעל ה-p-value הנמוך ביותר, מסף מסוים).
3. ממשיכים להוסיף משתנים עד שאחד מהתנאים הבאים מתקיים:
   - ה- p-value גבוהה מידי.
   - ביצועי המודל לא משתפרים באמצעות AIC, Adjusted R או BIC.

- Backward Elimination
1. מתחילים עם כל המשתנים העצמאיים.
2. מסירים את המשתנים הכי פחות משמעותיים עד שלא ניתן להסיר משתנים נוספים ללא פגיעה בביצועי המודל.

– Bidirectional Selection
מבצעים Forward ו- Backward Elimination באופן אופטימלי.

בעיתיות של stepwise –
- עלולה לגרום ל- overfit מאחר שבודקים המון מודלים, חישוב של $Adjusted\ R^2$ אופטימלי מידי. יש המון דרגות חופש בבחירת קריטריון.

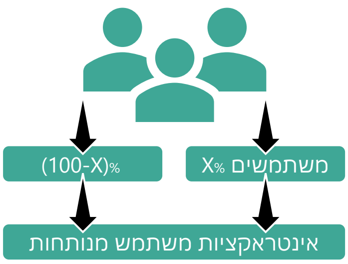
קריאת מודל רגרסיה –

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  GRADE   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.353
Method:                 Least Squares   F-statistic:                     6.646
Date:                Thu, 03 Oct 2024   Prob (F-statistic):            0.00157
Time:                        16:15:31   Log-Likelihood:                -12.978
No. Observations:                  32   AIC:                             33.96
Df Residuals:                      28   BIC:                             39.82
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
GPA            0.4639      0.162      2.864      0.008       0.132       0.796
TUCE           0.0105      0.019      0.539      0.594      -0.029       0.050
PSI            0.3786      0.139      2.720      0.011       0.093       0.664
const         -1.4980      0.524     -2.859      0.008      -2.571      -0.425
==============================================================================
Omnibus:                        0.176   Durbin-Watson:                   2.346
Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.167
Skew:                           0.141   Prob(JB):                        0.920
Kurtosis:                       2.786   Cond. No.                         176.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
```

- **Dep. Variable –** המשתנה התלוי Y – המשתנה שחוזים.
- **No. Observations –** מספר התצפיות בניסוי.
- **העמודה הראשונה בטבלה** מייצגת את המשתנים הבלתי תלויים X.
- **coef –**מציג את השפעת המקדם הבלתי תלוי במשוואת הרגרסיה.
- **Std err –** בודק כמה האומדן של המקדם מדויק, קטן יותר← יציב יותר.
- **Const –** השורה האחרונה בטבלה – הקבוע b.
- **P>t –** האם המשתנה מובהק סטטיסטית אם הערך קטן מ- a.

משוואת הישר תהיה הכפלת האיברים עם והוספת ה- const, כלומר: $y = \sum coef \cdot row_i + const$.
עבור הרגרסיה בתמונה נקבל – $y = 0.4639 \cdot GPA + 0.0105 \cdot TUCE + 0.376 \cdot PSI + (-1.498)$.

מבחני AB-

מבחני AB הם מבחנים הנשלטים בזמן אמת. זוהי דרך לקבל החלטות מבוססות נתונים, שיכולים לעזור לחזות מה יעבוד יותר טוב. וגם זהו מבחן שאפשרי בהרבה מקרים.



A/A Test

מטרת המבחן -

- בדיקת מערכת הניסוי
- חישוב השונות בסביבת הניסוי –
- לחשב רווחי סמך

לעיתים נקרא גם Null Test

המדגם

במבחנים נבדוק אם יש שינוי משמעותי סטטיסטית בין הממוצעים- $E(B) = \bar{x}_b - \bar{x}_a$ ונבדוק באיזו יחידה נבצע את הבדיקה, לפי מה נבצע רנדומליזציה. נרצה לבחור יחידה שנוכל לעקוב אחריה לפני ואחרי הניסוי.

בבחירת גודל המדגם נרצה לבחור פשרה בין גודל המדגם לזמן הניסוי, נרצה שהניסוי ירוץ מספיק זמן על מנת להפחית את ה- novelty effect ולהפחית אילוצים של זמנים (סופ"ש, שעות מאוחרות). על מנת להגיע לרווח סמך של 95% נדרוש ש- $n = \frac{16\sigma^2}{\Delta^2}$ כאשר σ^2 הוא השונות של ה- OEC וה- Δ הוא השינוי המינימלי שנרצה לזהות.

הגרלה הדרגתית -

האחוז צריך להתחיל באחוז נמוך של התעבורה, במידה ויש הצלחה אז נגדיל את האחוז בהדרגה. ההגדלה ההדרגתית נועדה למדוד הצלחה עם משאבים לא גבוהים, והורדת סיכונים.

השיקולים בבחירת המדידות הן לאסוף כמה שיותר מדידות על מנת לקבל הרבה נקודות מבט. הגדרת מדד עיקרי שיקרה OEC Overall Evaluation Criterion – עדיף שיהיה ארוך טווח, ולבצע בדיקות רבות.

סוגי מדדים –

- **מדדי יעדים – Goal Metrics –** המדדים שמעניינים את הארגון
- **מדדי ביניים – Driver metrics, indirect metrics –** מדדים קצרי טווח שלדעתנו יכולים לנבא את מדדי היעד של הניסוי.
- **מדדי אילוצים – Guardrail metrics –** מדדים לפעולות שלא נרצה שיקרו.

נרצה שהמדדים יהיו כמה שיותר פשוטים, יציבים, קשורים ליעד, ניתנים לפעולה, רגישים וחסינים מפני מניפולציות. כאשר מידה נהפכת למטרה היא מפסיקה להיות מדידה טובה Goodhart's law.

במהלך הניסוי –

- **אפקט הראשוניות –** כמה זמן לוקח למשתמשים להתרגל? עוזר מול קבוצת ביקורת.
- **אפקט החידוש –** משתמשים אוהבים לנסות דברים חדשים.
- **אפקט ההעברה –** שאריות ההשפעה מהמצב הקודם משפיעות על המצב החדש.

תופעות של תחילת הניסוי –

- רווחי הסמך קטנים ו- n גדל.
- התוצאות שואפות לממוצע.

הרצת ניסויים במקביל –

הרצה במקביל מאפשרת להקטנת הסיכוי להתנגשויות בכך שנבחר מבחנים על אזורים שונים, נחלק את המשתמשים לקבוצות זרות, נבדוק התנגשויות פוטנציאליות על קבוצות קטנות ונמדל באופן סטטיסטי את ההשפעות של השינויים.

אנליזה של תת אוכלוסיות –

נרצה לבחור תתי אוכלוסייה לפי מדדים שקשורים לניסוי, ונרצה להיזהר מפרדוקס סימפסון – תופעה שבה מגמה שמופיע במספר שונות נעלמת או מתהפכת כשנתונים מתאגדים ביחד, לכן צריך לשים לב על כל תת קבוצה.

טעויות נפוצות בניסוי A/B –

- חוסר בכוח סטטיסטי.
- בדיקת התוצאות לפני סיום הניסוי – עלולה להוביל לשינוי בביצוע הניסוי.
- הנחות לא נכונות על יציבות האוכלוסייה.
- הטיית הישרדות.
- הטיית כוונה לטיפול

בעיות בעקבות ניסוי A/B –

- עקיבה אחרי יחידות המדידה, השפעות חיצוניות.
- הסברים שלא מתייחסים לסיבתיות
- השפעות ארוכות טווח והשפעות חיצוניות.
- מחויבות ליישם את המצב החדש
- עקביות בחוויית המשתמש.

קושי להריץ ניסויים במקביל.

זה כל הקורס, בהצלחה 😊 אופק.

נספחים –

מדריך מחשבון –

אמור לעבוד למחשבונים הרגילים והטובים של קסיו, אני עם CASIO – 991ES PLUS, מאמין שהמקשים כמעט זהים, למחשבונים אחרים. כמעט הכל אותו הדבר.

הגדרה ראשונית

0. מומלץ לפני כל תרגיל לעשות איפוס למחשבון באמצעות SHIFT + 9 ואז מאשרים עם 3 ולוחצים = ועוד פעם = כעת המחשבון נקי.
1. לבחור MODE ואז 3 (סטטיסטי).
2. יש שני מצבים שכדאי לבחור מהם –
   - 1 - **חד ממדי** - קריטריון אחד נוח לרוב הדברים בקורס.
   - 2 - **דו ממדי** - שימושי בעבור רגרסיה ו- ANOVA דו צדדית.
3. כעת תהיה לכם טבלה עם הנתונים מאשרים באמצעות =, ניתן לנווט עם החיצים ◀▶▲▼.
4. כשסיימתם ללחוץ על AC לאישור הטבלה ונעבור לצורה שנראית כמו המחשבון הרגיל כשיש כיתוב למעלה STAT כל האפשרויות שנראה הן בזכות שאנחנו במצב STAT.
5. במידה ונרצה לחזור למצב רגיל נבחר - לבחור MODE ואז 1 (COMP).

אפשרויות סטטיסטיות –

לחיצה על SHIFT + 1 תביא אותנו לבקרה על הסטטיסטיקה. יש כמה אפשרויות:

- 1 - **Type** - עריכת הסוג בדיקה – יחזיר אותנו למסך שראינו ב- 2, אם משנים.
- 2 - **Data** - עריכת הנתונים שהכנסנו.
- 3 - **Sum** - סכומים, יהיו לנו שם נתונים שימושיים כמו $\sum x$ ו-$\sum x^2$.
- 4 - **Var** - נתונים סטטיסטים, נוכל למצוא שם את n מספר הנתונים, μ_x – תוחלת, σ_x – סטיית תקן באוכלוסייה. ו- S_x – סטיית תקן במדגם
- 5 - **Dist** - (בחד ממדי) התפלגויות כמו Z ו- t, לא נוח במיוחד לשימוש, הטבלה עדיפה.
- 5 - **Reg** - (בדו ממדי) רגרסיות - יש לנו את מה שנרצה עבור רגרסיה $R = r$, ועבור רגרסיה ליניארית נוכל להגדיר קו מגמה באמצעות A – הקבוע ו-B המקדם של X.
- 6 - **MinMax** - הערכים המינימליים והמקסימליים של X .

בעבור התפלגויות דו ממדיות יהיו לנו אפשרויות גם ל- x וגם ל- y ב- 3, 4 ו- 6.

מטריצות במחשבון –

1. לבחור MODE ואז 6 (מטריצות).
2. יש ברשותנו 3 מטריצות שנוכל לשחק איתן נבחר את אחת מהן.
3. נבחר את גודל המטריצה (יש עד 3 × 3).
4. נכתוב את המטריצה, מאשרים באמצעות =, ניתן לנווט עם החיצים ◀▶▲▼.
5. ניתן לזמן את המטריצה על ידי לחיצה על SHIFT + 4.

- ניתן לקרוא למטריצות שהזנו ולעשות עליהן פעולות כמו חיבור, כפל והופכית (עם x^{-1}).
- פונקציית transpose באמצעות SHIFT + 4, עם 6 ובתוך הסוגריים נקרא למטריצה.
- ניתן לשנות את המטריצות באמצעות SHIFT + 4, עם 1 לבחירת הגודל ו- 2 לשינוי הנתונים.

הרחבת דף הנוסחאות למילואימניקים WIP

- **תוחלת הסטטיסטי שווה לממוצע המדגם -** $E[\bar{X}] = E[X]$ וסטיית התקן היא - $\sigma(\bar{X}) = \frac{\sigma(x)}{\sqrt{n}}$
- **חישוב אמד מבחן -** $\frac{E[X]-\mu_0}{S_x/\sqrt{n}}$
- **סכום סדרה חשבונית -** $\frac{n(2a_1+(n-1)d)}{2}$
- **יחסי התפלגות -** $\frac{N(0,1)}{\chi^2(n)/n} = t(n)$
- **תיקון Šidák –** תיקון פחות מחמיר לתיקון רמת המובהקות כאשר מבוצעות בדיקות מרובות, נגדיר את a להיות - $a_{Šidák} = 1 -$ $(1-a)^{\frac{1}{n}}$ משמש להפחת הסיכוי לטעות טעות מסוג 1.
- **מציאה מהטבלה בהתפלגות נורמלית –** $\Phi = P(Z_X \le Z_y) = P(X \le y)$ כאשר X זה המשתנה המתפלג נורמלי, Z_x, Z_y אלו הערכים המתוקננים, ו- Φ מייצג את הערכים בטבלה, עבור $n > 30$ נחפש בטבלת Z, אחרת בטבלת t.

שלבים לביצוע מבחן One Way ANOVA –

- $SS_{Total} = \sum_{j=1}^{q}[(n_j-1)s_j^2 + n_j(\bar{y}_j - \bar{\bar{y}})^2]$ – סכום הריבועים הכולל של כל האיברים . $df_{Between} = q-1$
- $SS_{Within} = \sum_{j=1}^{q}\sum_{i=1}^{n_q}(y_{i,j} - \bar{y}_j)^2 = \sum_{j=1}^{q}(n_j-1)s_j^2$ – סכום ההשתנות בתוך הקבוצות. $df_{Within} = n-q$, $MS_W = \frac{SS_{Within}}{df_{Within}}$
- $SS_{Between} = \sum_{j=1}^{q} n_j(\bar{y}_j - \bar{\bar{y}})^2$ – ההשתנות בין הממוצעים של הקבוצות השונות. $df_{Total} = n-1$, $MS_B = \frac{SS_{Between}}{df_{Between}}$

1. חישוב ממוצע לכל קבוצה.
2. נחשב ממוצע של כל הקבוצות ביחד.
3. נמצא את סכום השונויות בתוך כל הקבוצות - $SS_{Within} = \sum_{j=1}^{q}\sum_{i=1}^{n_q}(y_{i,j} - \bar{y}_j)^2$.
4. נמצא את סכום השונויות בין הקבוצות - $SS_{Between} = \sum_{j=1}^{q} n_j(\bar{y}_j - \bar{y})^2$.
5. נמצא את דרגות החופש - $df_{Between} = q-1,\ df_{Within} = n-q, df_{Total} = n-1$.
6. למצוא את הממוצע שונות בין הקבוצות MS_B עבור $SS_{Between}$. $MS_B = \frac{SS_{Between}}{df_{Between}}$.
7. חישוב ממוצע השונות בתוך הקבוצות MS_W עבור SS_{Within}. $MS_W = \frac{SS_{Within}}{df_{Within}}$.
8. נחשב את ה- $F_{Statistic}$ סטטיסטי מדגם - $F_{Statistic} = \frac{MS_B}{MS_W}$.
9. חישוב F_{Critic} מהטבלת F באופן הבא- $F_a\left(\overbrace{q-1}^{df_b}, \underbrace{n-q}_{df_w}\right)$.
10. אם $F_{Statistic}$- קטן מ- F_{Critic} נדחה את H_0.

שלבים לביצוע מבחן Two Way ANOVA –

בהינתן שני משתנים בת"ל- עבור משתנה A בגודל p ומשתנה B בגודל q.

1. נגדיר מערכת השוואות לשני משתנים בנפרד ומערכת אחת על השפעה בין שני הגורמים.
2. נחשב ממוצעים עבור כל משתנה ועבור אינטראקציית משתנים. מומלץ לארגן בטבלה.

חישוב כולל -

3. נחשב ממוצע לכל הערכים $\bar{y}_{total}$.
4. נחשב את SS_{Total} כאשר $SS_{Total} = \sum_{m=1}^{p}\sum_{i=1}^{q}\sum_{j=1}^{n_{i,j}}(x_{m,i,j} - \bar{y}_{total})^2$.
5. חישוב דרגות החופש עבור B ,A והכולל - $df_A = p-1$, $df_B = q-1$, $df_{Total} = n \cdot p \cdot q - 1$.
6. חישוב $MS_T = \frac{SS_{Total}}{df_{Total}}$

חישוב עבור A ו- B -

7. חישוב ה- SS_A וה- SS_B - $SS_A = \sum_{i=1}^{p} n_i(\bar{A}_i - \bar{y}_{total})^2$, $SS_B = \sum_{i=1}^{p} n_i(\bar{A}_i - \bar{y}_{total})^2$
8. חישוב MS_A וה- MS_B - $MS_A = \frac{SS_A}{df_A}$ ו- $MS_B = \frac{SS_B}{df_B}$

עבור בתוך D- Between AB

9. חישוב ה- SS_{AB} - $SS_{AB} = SS_{Between} = \sum_{i=1}^{p}\sum_{j=1}^{q}(\overline{A_iB_j} - \bar{A}_i - \bar{B}_i + \bar{y}_{total})^2$
10. חישוב df_{AB} - $df_{AB} = (p-1)(q-1)$
11. חישוב MS_{AB} - $MS_{AB} = \frac{SS_{AB}}{df_{AB}}$

עבור השגיאה Within AB.

12. חישוב ה- SS_{Within} - $SS_{Within} = \sum_{m=1}^{p}\sum_{i=1}^{q}\sum_{j=1}^{n_{i,j}}(x_{m,i,j} - \overline{AB}_{i,j})^2$
13. חישוב df_{Within} - $df_{Within} = (n-1) \cdot p \cdot q$
14. חישוב MS_{Within} - $MS_{Within} = \frac{SS_{Within}}{df_{Within}}$

חישוב F וקבלה או שלילה -

15. נחשב את F_A, F_B ו- F_{AB} - $F_A = \frac{MS_A}{MS_{Within}}, F_B = \frac{MS_B}{MS_{Within}}, F_{AB} = \frac{MS_{AB}}{MS_{Within}}$
16. נחשב F_{critic} עבור F_A, F_B ו- F_{AB} באופן זהה: - $F_x(df_x, df_{Within})$
17. אם $F_{Statistic}$- קטן מ- F_{Critic} נדחה את H_0.
18. במידה ודחינו את H_0, נרצה לדעת איזה קבוצה שונה מהאחרות, נשתמש במבחן post hoc.

- בהתפלגות נורמלית ממוצע המדגם שווה לתוחלת הסטטיסטי כך : $E[X] = E[\overline{X}]$ וגם מ תקיים $\sigma(X) = \sigma(X)/\sqrt{n}$
- זהות חשובה- $\mathrm{COV}(X, Y) = (\mathrm{var}(x) + \mathrm{var}(Y) - \mathrm{var}(X + Y))/2$

רגרסיות -

- $\boldsymbol{SS_{explaned}}$ – החלק של השונות שתלוי ב- X. $SS_{Reg} = S_{explaned} = \sum(\hat{y}_i - \bar{y})^2$
- $\boldsymbol{SS_{residual}}$ – החלק בשונות שלא תלוי ב- X, המינימום שגיאות לקו. $S_{residual} = \sum(y_i - \hat{y}_i)^2$
- $\boldsymbol{SS_{Total}}$ – סכום הריבועים של השונות של הנתונים המקוריים. $S_{Total} = \sum(y_i - \bar{y})^2$
- נקבל גם ש- $R^2 = \frac{SS_{explaned}}{SS_{Total}}$
- לכן, נתייחס ל- R^2 בתור אחוז השבר של השונות המוסברת ע"י הרגרסיה. נשים לב גם שמתקיים - $S_{Total} = S_{explaned} + S_{residual}$ לכן ניתן לכתוב גם - $R^2 = \frac{SS_{Total} - S_{residual}}{SS_{Total}}$.

טבלת הסתברויות, תוחלות ושונויות

סוג המשתנה המקרי	פונקציית ההתפלגות	תוחלת	שונות	אמד נראות מרבית – אנ"מ	חסר-הטיה?
התפלגות ברנולי $\boldsymbol{X \sim Ber(p)}$	$P(X = x) = \begin{cases} 1 & x = 1 \\ 1 - p & x = 0 \end{cases}$	p	$p(1 - p)$	$= \frac{X}{n}$	כן
התפלגות בינומית $\boldsymbol{X \sim Bin(n, p)}$	$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$	np	$np(1 - p)$	$= \frac{X}{n}$	כן
התפלגות גיאומטרית $\boldsymbol{X \sim Geo(p)}$	$P(X = k) = (1 - p)^{k-1} \cdot p$	$\frac{1}{p}$	$\frac{1 - p}{p^2}$	$= \frac{1}{\bar{X}}$	לא (כלפי מטה)
התפלגות אחידה בדידה $\boldsymbol{X \sim U(a, b)}$	$P(X = k) = \frac{1}{b - a + 1}, k \in (a, a + 1 \ldots b)$	$\frac{a + b}{2}$	$\frac{1 - (b - a + 1)^2}{12}$	$= max\{X1 \ldots Xn\}$	לא (כלפי מטה)
התפלגות מעריכית $\boldsymbol{Exp(\lambda)}$	$P(x, \lambda) = \lambda e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	עבור $\boldsymbol{\lambda}$: $= \frac{1}{\bar{X}}$ עבור μ : $= \frac{1}{\bar{X}}$	לא (כלפי מעלה) כן
התפלגות פואסון $\boldsymbol{X \sim Pois(\lambda)}$	$P(X = k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$	λ	λ	$= \bar{X}$	כן
נורמלית	$\frac{1}{\sqrt{2\pi\sigma^2}}{}^{-(\frac{(x-\mu)^2}{2\sigma^2})}$	μ	σ	תוחלת: $= \bar{X}$ שונות $= \frac{\sum_i (X_i - \bar{X})^2}{n}$	כן לא (כלפי מטה)

גודל מדגם	התפלגות	שונות	הטווה:
$\boldsymbol{n > 30}$	התפלגות דגימה נורמלית	ידועה	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(\mu, \frac{\sigma^2}{n})$
	התפלגות דגימה נורמלית	לא ידועה	$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim N(\mu, \frac{\hat{S}}{n})$
$\boldsymbol{n \leq 30}$	דגימה מאוכלוסייה מתפלגת נורמלית	ידועה	$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(\mu, \frac{\sigma^2}{n})$
	דגימה מאוכלוסייה מתפלגת נורמלית	לא ידועה	$\frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim t(n - 1)$

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.

שם המבחן	סוג המבחן	הנחות	מתי משתמשים?	דוגמה לנתונים
One-Way ANOVA	פרמטרי	נורמליות, שוויון שונות	השוואת ממוצעים בין יותר משתי קבוצות לקבוצה אחת	השוואת ציוני תלמידים בשלושה בתי ספר שונים
Two-Way ANOVA	פרמטרי	נורמליות, שוויון שונות	בדיקת השפעה של **שני משתנים** בלתי תלויים על משתנה תלוי	השפעת סוג הדיאטה והפעילות הגופנית על ירידה במשקל
Mann-Whitney U	א-פרמטרי	אין הנחת נורמליות	השוואת שתי קבוצות בלתי תלויות כאשר הנתונים לא מתפלגים נורמלית	השוואת גובה בין גברים לנשים
Wilcoxon Signed-Rank Test	א-פרמטרי	אין הנחת נורמליות	השוואת מדדים באותו מדגם (מדגם מזווג)	בדיקת שיפור בזיכרון לפני ואחרי קורס אימון מוחי
Kruskal-Wallis	א-פרמטרי	אין הנחת נורמליות	השוואת ממוצעים של יותר משתי קבוצות כאשר הנתונים אינם נורמליים	השוואת ציוני מבחן בין שלוש אוניברסיטאות שונות
Chi-Square Test (מבחן כי-בריבוע לאי תלות)	א-פרמטרי	משתנים קטגוריים בלבד	בדיקת תלות בין **שני** משתנים קטגוריים	בדיקה אם יש קשר בין עישון לסוג מחלה
Chi-Square Goodness-of-Fit (מבחן טיב ההתאמה)	א-פרמטרי	משתנים קטגוריים בלבד	בדיקה אם ההתפלגות של משתנה **אחד** תואמת התפלגות תאורטית צפויה	בדיקה אם חלוקת הקולות בבחירות תואמת את הציפייה
Tukey's HSD	פוסט-הוק	נורמליות, שוויון שונות	השוואת קבוצות לאחר ANOVA	בדיקת אילו קבוצות שונות משמעותית זו מזו לאחר ANOVA

סיכומון מבחני AB – לדף נוסחאות

מבחני **A/B** הם ניסויים מבוקרים בזמן אמת המשמשים להשוואת שתי גרסאות של מוצר, ממשק או אתר, במטרה לזהות איזו גרסה משיגה ביצועים טובים יותר. לפני הרצת A/B, מבצעים **A/A Test** כדי לוודא שהמערכת יציבה ולחשב שונות ורווחי סמך. בניסוי, מחלקים את המשתמשים **באופן רנדומלי** לקבוצות, כאשר המדגם חייב להיות גדול מספיק כדי להגיע **למובהקות סטטיסטית**, תוך איזון בין גודל המדגם למשך הניסוי.

כדי לנתח את התוצאות, משתמשים **במדדי יעד (Goal Metrics)** למדידת ההצלחה, **מדדי ביניים (Driver Metrics)** לניבוי השפעות, ו**מדדי אילוצים (Guardrail Metrics)** למניעת פגיעה בחוויית המשתמש. במהלך הניסוי חשוב להיזהר מ**הטיות** כמו **אפקט חדשנות, השפעות חיצוניות, והטיית הישרדות**. כמו כן, הרצת ניסויים במקביל דורשת ניהול קפדני כדי למנוע **התנגשויות** בין קבוצות שונות. בסוף הניסוי, רק תוצאות **מובהקות סטטיסטית** מספקות תובנות אמינות לשיפור המוצר.

מחוץ לדף הנוסחאות

הוכחות שכדאי לזכור –

צ"ל: $MSE(\hat{\theta},\theta) = E\left[(\hat{\theta}-\theta)^2\right] = Var_\theta(\hat{\theta}) + Bias_\theta(\hat{\theta},\theta)^2$

הוכחה:

$$
\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_\theta\left[(\hat{\theta}-\theta)^2\right] \\
&= \mathrm{E}_\theta\left[\left(\hat{\theta}-\underbrace{\mathrm{E}_\theta[\hat{\theta}]+\mathrm{E}_\theta[\hat{\theta}]}_{\text{טריקים שפלים}}-\theta\right)^2\right] \\
&= \mathrm{E}_\theta\left[\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)^2+2\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)+\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathrm{E}_\theta\left[\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)^2\right]+\mathrm{E}_\theta\left[2\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)\right]+\mathrm{E}_\theta\left[\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)^2\right] \\
&= \mathrm{E}_\theta\left[\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)^2\right]+2\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)\mathrm{E}_\theta\left[\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right]+\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)^2 && \mathrm{E}_\theta[\hat{\theta}]-\theta=\text{const.} \\
&= \mathrm{E}_\theta\left[\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)^2\right]+2\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)\left(\mathrm{E}_\theta[\hat{\theta}]-\mathrm{E}_\theta[\hat{\theta}]\right)+\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)^2 && \mathrm{E}_\theta[\hat{\theta}]=\text{const.} \\
&= \mathrm{E}_\theta\left[\left(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]\right)^2\right]+\left(\mathrm{E}_\theta[\hat{\theta}]-\theta\right)^2 \\
&= \mathrm{Var}_\theta(\hat{\theta})+\mathrm{Bias}_\theta(\hat{\theta},\theta)^2
\end{aligned}
$$

צ"ל: $Y = X\cdot W$ אז $W = (X^TX)^{-1}X^TY$

הוכחה: אם A מטריצה סימטרית מתקיים $\frac{\partial}{\partial z}(z^TAz) = 2Az$ מאחר ש- $z^TAz = \sum_{i=1}^n\sum_{j=1}^n z_iA_{ij}z_j$,.
הנגזרת לפי z: $\underbrace{\frac{\partial(z^TAz)}{\partial z_k} = \sum_{i=1}^n z_iA_{ik} + \sum_{j=1}^n A_{kj}z_j}_{\text{הם שווים}}$. אז אם נרצה למזער את $E = (Y-XW)^T(Y-XW)$ נפתח ונקבל
ש- $E = Y^TY - Y^TXW - W^TX^TY + W^TX^TXW = Y^TY - 2Y^TXW + W^TX^TXW$ נגזור ונשווה לאפס ונקבל-
$\Rightarrow \frac{dE}{dW} = 2Y^TX - 2W^TX^TX = 0 \Rightarrow X^TXW = X^TY$ אז: $(X^TX)^{-1}X^TY = (X^TX)^{-1}X^TXW \Rightarrow$ סך הכל יצא- $\boldsymbol{W = (X^TX)^{-1}X^TY}$

צ"ל: $R^2 = \frac{SS_{explaned}}{SS_{Total}}$

הוכחה: $R^2 = r^2 = \frac{\sigma_{xy}^2}{\sigma_x^2\sigma_y^2} = \frac{\overbrace{w^2}^{w=\frac{\sigma_{xy}}{\sigma_x^2}}\ \sigma_x^2}{\sigma_y^2} = \frac{\overbrace{w^2E(x-\bar{x})^2}^{var(ax)=a^2\cdot var(x)}}{\sigma_y^2} = \frac{\overbrace{E((b+wx-\bar{y})^2)}^{\bar{y}=w\bar{x}+b}}{\sigma_y^2} = \frac{\overbrace{E(\hat{y}-\bar{y})^2}^{\hat{y}=wx+b}}{E(y-\bar{y})^2} = \frac{SS_{explaned}}{SS_{Total}}$