מבוא לסטטיסטיקה

קוזלובסקי

הרצאה מס' 1 - 30/10/00

מושגי יסוד

Variable (משתנה) - ממד/תכונה בעלת אפשרויות התבטאות רבות. למשל גלאים.(מכיל טווח גדול של אפשריות).

Constant (קבוע) - נתון קבוע שאינו משתנה. היות וכך, לרוב הוא פחות מעניין.

שני הנ"ל יכולים להתחלף במשמעויותיהם בהתאם למצבי הרקע הקיים. למשל ממדי חיים ומוות בכיתת לימוד מהווים מצב קבוע אך בבי"ח הם מהווים משתנה היות ומס' המתים/חיים שם הוא משתנה.

Continuous (רציף) - מספר אשר רמת הדיוק שלו היא אין סופית. למשל גובה של בן-אדם או המשקל שלו.

Discrete (בדיד) - נתון מוחלט. למשל מספר פריטים באוכלוסייה.

Infinite

Finite

Sample (מדגם) - על פי פרמטרים מסוימים.

Population (אוכלוסייה) - מספר פריטים בעלי תכונות זהות. מסיק מהמדגם על האוכלוסייה

Independent variable - משתנה בלתי תלוי.

Dependent variable - משתנה תלוי. $Y=f(x)$ לעתים קיים קושי לזהות את המשתנה התלוי והבלתי תלוי.

אם יש פער בזמן בין המשתנה הבלתי לתלוי ניתן להגדיר ביתר קלות מי הם המשתנים. באופן מקביל לסיבה ותוצאה, מהוא הפער בין השניים, מה הקשר ביניהם. במידה וקיים ספק בזיהוי אופי המשתנים, ניתן להתייחס אליהם כאל משתנים בלבד. (דוגמת חברת יצרניות הסיגריות - עישון לא מוביל לסרטן. קיים משתנה נוסף אשר הוא הגורם למחלה ויגרום לה בין אם האדם יעשן או לא).

הרצאה 2 - 6/11/00

סולמות - Scales

Nominal (נומינאלי) - משתנה. למשל מין (gender) אמנם אין יותר מדי אפשרויות אבל ציון של משתנה זה המדגם לא מציין או מהווה העדפה של הנתון לכיוון זה או אחר. הוא אפיון קבוצתי של קטגוריה תיאורית בלבד. אין היררכיה כזו או אחרת. מספר ת.ז. הוא שמי. אין בו שימוש יום יומי אבל יחד עם זאת ניתן לקטרג אותו. המיון הוא לא מדור, אין משמעות לסדר.

ordinal (אורדינאלי/דירוגי) - יש סדר בהגדרות. מאפיין בסדר אורדינאלי הוא שיש מדרג אך ההפרש בין דרג אחד לשני אינו ידוע/לא קיים מבחינת המדע. (לדוגמה - דירוג מלכות יופי או העדפה למשקה). אין דרך להגיע למשמעות של ההפרש.

Interval (אינטרואלי) - בין אורדינלי ליחסי. קיים מדרג אבל לא בייחוס לאפס מוחלט משום שהוא בעייתי להגדרה. קיימת משמעות להפרשים בין המדדים ולהשוואה בין פערים על אותה סקלה. נקודת היחס היא שרירותית על הסקלה. למשל - ציונים במבחן מול רמת הידע של הנבחן. מה רמת הידע שמגלם הציון? האם מי שקיבל 90 יודע פי 2 ממי שקיבל 45? האם קיימת משמעות להפרש שבין שני הציונים?

ratio (יחס/רציונלי) - כמו בפיסיקה, היחס בין מדדים הוא בעל משמעות כמותית. הוא בהשוואה לאפס המוחלט (absolute zero) המהווה את נקודת היחס, בניגוד למספרים הסידוריים בסולם האורדינלי או למספרים הסמליים בסולם הנומינלי.

התפלגות

דוגמת IQ

מאגר גדול של נתונים ניתן להגדיר ע"פ סולם של מדרוג , רמת השכיחות (f) המופיעה במדרוג היא ההתפלגות.ב שימוש במדרוג / אינטרבלים הופך את תוצאות ההתפלגות לנכונות יותר. מה גם שעיבוד נתונים להתפלגות הוא נוח יותר לעבודה. השאיפה היא לכל היותר 20 אינטרבלים ובעדיפות אפילו לא יותר מ 10. טווח האינטרבלים תלוי באופי נתונים ובטווח הבסיסי שלהם. טווחים מקובלים הם 5, 10, 50 או 100. התחלת הסדרה תהיה במספר שמתחלק בגודל האינטרבל ובספרה הגדולה יותר:

Class interval	
התפלגות	f
125-129	0
120-124	2
115-119	2
110-114	1
105-109	1

אינטרבל אמיתי / Real Exact Interval - אינו משאיר פערים בין שלב ושלב באנטרבלים. מתמודד בעיקר עם מדדים רציפים. במקרה כזה האינטרבלים יהיו :

124.5-129.5
124.5 -120.5

מאחר ואנו מוגבלים למכשירי מדידה אנו משתמשים ב class interval. סביר להניח שמדדים שאנו מתעסקים איתם הם רציפים אבל אין לנו דרך אמינה לבדוק זאת.

Cumulative Frequency

שכיחות מצטברת מציגה את סה"כ השכיחות:

f	Cf
0	6
2	6
2	4
1	2
1	1

הרצאה 3 - 13/11/00

צורת גרף התפלגות (המקרה זה של משתנה איכותי בדיד): העמודה בנויה על mid point סכום הגבול האמיתי העליון עם הסכום האמיתי התחתון

חלקי 2. ע"פ הספר יש גבול בין כל עמודה, ע"פ קוזלובסקי הרווח הוא לא נכון.

סיגמה Σ

סיגמה היא סכום של אברים בתחום מסוים. למשל:

Σxi=X1+X2+X3...+Xn

כשהטווח הוא i=1, i=n

n מסמלת את המשתנה האחרון.

כל אחד מהמשתנים הם בלתי תלויים. מיקומם ברשימה הוא לא קבוע או דירוגי. כלומר אם רוצים לעשות Σ של 1X ושל 3X משנים את מיקומם ברשימה כך שיהיו עוקבים, 1 ו 2.

חוקים

חוק 1 - ישנם משתנים קבועים. C מסמן קבוע ו N הוא מספר החזרות של הקבוע

i=N

$$\sum_{i=1} c = Nc$$

אם c שווה 4 ו N שווה 5 אז התוצאה היא 20.

חוק 2 - נתן להוציא את הקבוע אל מחוץ לΣ

$$\sum cXi = c\sum Xi$$

ממוצע $\bar{X}$

השימוש הוא במילה mean

Mean = ΣXi/N

m הוא ממוצע של אוכלוסייה.

הרצאה 4 - 7/11/00

כל התפלגות נורמלית בנויה בצורה דומה, ללא שום קשר לדבר הנמדד.

נוסחתו של גאוס - Se-x2fx

Variance 1 - שונות (של מדגם): S2 = Σ(x-x)2/N-1

Standard deviation - (סטית התקן): שורש של S2

Variance 2 שונות (של אוכלוסיה) - σ2 = Σ(x - x)2 / N

היחס בין השונות הוא ש expected של s2 הוא σ2

כשאר אנו עושם דגימה מתוך אוכלוסיה עלינו להשתמש רק במונה של N-1, רק כך ממוצע השונויות יהיה שווה לשונויות האוכלוסיה.
אם משהו הוא קבוע מבחינת הממוצע (הממוצע הוא ידוע ומחייב) הוא "כופה" את נתוני ההתפלגות. חלק מהנתונים הוא ידוע.
בהתפלגות נורמלית, סטית תקן אחת מעל/ מתחת לממוצע כוללת 34% מהאוכלוסיה בתחום שהיא תוחמת. סטית התקן השניה מכילה כ 13%. כך בהתפלגות נורמלית של כל דבר.

הרצאה 6 - 11/12/00

הסתברות

Mutually exclusive, מאורעות חריגים.

P(AUB)=P(A) + P(B) - P(A∩B) הסתברות האיחוד של A ו B היא ההסתברות של A ועוד ההסתברות של B פחות החיתוך של ההסתברות של A ו B.

union - U, איחוד.

∩ - intersection, overlap, חיתוך.

למשל סטודנט הלומד בבר אילן וסטודנט הלומד בטכניון. הסיכוי שהוא יהיה גם פה וגם שם הם אפסיים (להלן מאורע חריג), אין חיתוך בין המיקרים ולכן P(A∩B) שווה לאפס.

Independent events, מאורעות בלתי תלויים.

2 דגמאות הסתברותיות שאפשריות או לא אפשריות ביחד.

(P(A∩B) = P(A) * P(B

למשל :

P(A) = .12, תלמידים בבר אילן.

P(B) = .05, אכילת פלאפל.

P(A∩B) = .006

P(AUB) = P(A) + P(B) - .006 = .12 + .11 - .006 = .164, ההסתברות להיות בבר אילן או לאכול פלאפל או לעשות את שניהם.

Dependent event, מאורע תלוי- יש השפעה של נתון אחד על השני.

במקרה זה החיתוך אינו ידוע ויש לערוך תצפיות/ מחקר כדי לגלות מהו.

למשל:

P(A) = .20 ארועי גשם בשנה.

P(B) = .15 שימוש במטריה בשנה.

אם החיתוך היה שווה 11. אז התוצאה היא24..

- מקסימום החפיפה/ חיתוך האפשרית היא 15. מאחר ולא יכול להיות חיתוך גדול יותר מהערך הקטן מבין ההסתברויות.
- מקסימום האיחוד הוא קרוב ל .35 מאחר ואם היה בדיוק .35 אז הסוג היה אירוע חריג.

Permutation (סידורים?)

הפרמוטציה של n איברים הוא n פקטוריאלי. (כמו n עצרת)

$n! = (n-1)*(n-2)*(n-3)*...1$

הרצאה 7 - 18/12/00

הנוסחה: $n! / (n-r)!$

למשל סדרה של מספרים: A, B , C. אם ה n שווה ל 3 וה r שווה ל 2 אנו מדברים רק על קומבינציה של 2 מספרים. הסדר כאן חשוב! האפשרויות הן: AB, AC, BA, BC, CA, CB. באותה סידרה, ה n שווה עדיין ל 3 (כמות האברים) וה r הפעם שוה ל 3. הסדרות הן: ABC, ACB, BAC, BCA, CAB, CBA.

Combination צירופים

כאן הסדר לא חשוב.

הנוסחה: $n! / r! (n - r)!$

אם ה r וה n שווים יש רק אפשרות אחת.

בשלוש זריקות של מטבע האפשרויות הן:

כלומר 8 אפשרויות

מה ההסתברות לקבל רק פעמיים פ? - 3/8

מה ההסתברות לקבל לפחות פעמיים פ? - 4/8, או ½

מה ההסתברות לא לקבל בכלל פ? - 1/8 .

פפפ	ععע
פפע	ععפ
פעפ	עפע
פעע	עפפ

בינום binomial (של ברנולי).

ההסתברות לקבל = p.

הנוסחה: $(n r) p r (1 - P) n - r$

להוציא מלך אחד מחבילה כשכול פעם אני מחזיר את הקלף לחבילה:

n = 2 מספר המטלות/ מספר הניסויים

r = 1 מספר ההצלחות

P = 4/54.

(2 1) (4/54) (48/52) = 24/169

משולש פסקל כאשר p = q = 1/2 עם מטבע.

זריקות					1				
1				1		1			
2			1		2		1		
4		1		3		3		1	
5	1		4		6		4		1

הרצאה 11 - 8/1/01

תיאורית בייס Baye`s theory

הסתברות מותנית - $P(A1 / B1) = P(A1 \cap B1) / P(B1)$

		A1	A2	A3	
		פרטי	ציבורי	ברגל	
0.6	0.3	0.1	0.2	זכר	B1
0.4	0.25	0.5	0.1	נקבה	B2

) Pפרטי/זכר) = P(פרטי∩זכר) / P(זכר) 0.3 / 0.6 = 0.5

השלמת נתוני טבלה תלויה בידיעה שהקטגוריות הקיימות הן כול הקטגוריות שיש. כלומר הסה"כ הוא ידוע (או 1) ואז אפשר להשלים את הטבלה.

בנוסף אפשר לדעת ע"פ הנתונים בטבלה מה הסיכויים של חיתוך נתונים הטבלה להתקיים, ע"פ אינפורמציה חלקית בטבלה.

בינום : מה הממוצע של פלי מ 100 זריקות של מטבע? = 50

$M = n*p$ $100*0.5=50$

מה ה SD של פלי ב 100 זריקות?

$S = n*p*q$ שורש של $100*0.5*0.5 = 5$

שונות בבינום היא אותה נוסחה כמו סטית התקן רק ללא השורש.

לדוגמא : לקבל לפחות 7 פעמים פלי בתוך 10 הטלות מטבע.

ממוצע = $n*p = 0.5*10 = 5$

סטית התקן =שורש של $n*p*q$ = שורש של 10 * ½ * ½ = 1.58

בינום = C(10 7) * 0.57 * 0.53 + C(10 8) * 0.58 * 0.52 + C(10 9) * 0.59 * 0.51 + C(10 10) * 0.510 * 0.50 = 0.172

ככול שה n שואף לאין סוף, ההתפלגות שואפת להיות נורמלית וכך ניתן להשתמש בהתפלגות זו כדי לקבל את הסיכויים לתוצאה.

הבעיה היא שהנתון 7 הוא בדיד והגרף הוא רציף. לכן ניקח את הנתון המקורב - 6.5

x - x^ / SD = 6.5 - 5 / 1.58 = 0.949 = z

5.5 - 5 /1.58

כניסה עם ה 2 לגרף נותנת את התוצאה 0.171 שההיא מקורבת למדי בכדי להסתמך עליה כעל סיכויים. הקירוב יתאפס כשה n יהיה שווה לאין סוף.

תוצאה מדוייקת יכולה להתקבל רק בחישוב בינום.

הרצאה 12 - 15/1/1

מתאם קורלציה

טווח ה r נע בין 1 ל -1

נוסחת הקורלציה r = CoV (x , y) / √ S x 2 * S y 2

דוגמא 1 :

נתונים סדרת נתונים :

x	y
8	600
9	700
10	800

ממוצעים : x ^ = 9

y ^ = 800

חישובים ע"פ הנוסחה, המונה :

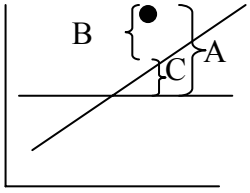
-) * (-1)	600 -	8 - 9
100)	700	
0	700 -	9 - 9
	700	
100) * (1)	800 -	10 - 9
(	700	
200		

המכנה : 200 / √ 2 * 20,000

התוצאה היא 1 = 200 / 200 כלומר מתאם חיובי מלא.

דוגמא נוספת בה המתאם r = .40, ממנה נוציא את r 2 = .16 או 16%. נתון זה אומר ש 16% מהשונות מוסבר ע"י המתאם. r 2 הוא השונות המוסברת.

לדוגמא : 4 נבדקים מסודרים על גרף. ביניהם נמתח קו אשר נחשב לטוב ביותר, לכול הנבדקים יש את המרחק המינימלי מהקו.



הנקודה היא y` היא הקשר הקיים בין X ל Y. היא מתארת הכי טוב את הנקודות. הקו האופקי הוא ממוצע Y.

r 2 השונות המוסברת של X , Y עם הקשר ביניהם. שונות של Y המוסברת ע"י X. לרוב Yהוא המשתנה התלוי (הביטוי העתיד לבוא).

B הוא המרחק של הנקודה מהקו, היא מראה את ניבוי הניסוי. ככול שהנקודה קרובה יותר לקו יש דיוק גדול יותר. אם הכול נמצא על הקו, השונות מוסברת באופן מושלם. כך יתאפשר ניבוי טוב יותר. כשמרחק זה יהיה שווה לאפס, הנקודה היא על הקו.

A	B	C
Y i - Y =	(Y I - Y`) +	(Y` - Y^)
ממוצע של שונות	מנובא	מנובא

A מיצג את השונות, התפזרות תוצאות המחקר.

spearman / Rank correlation

שימוש בנוסחה זו יעשה בסולמות מסוג סדר. P = 1-(6 Σ d 2 / N 3 - N)

למשל תחרות יופי:

מועמדת	שופט 1	שופט 2	d הפרש	d 2
A	3	2	1	1
B	1	1	0	0
C	2	3	-1	1
D	4	4	0	0
			Σ d 2 =	
			2	

כדי לדעת מה המתאם בין הדירוגים הנ"ל יש להשתמש במתאם ספירמן.

N=4

1 - [6 * (2) / 60] = 1 - 0.2 = 0.8

קיים קשר גבוה בין דירוגי השופטים.

הרצאה 13 - 26/2/01

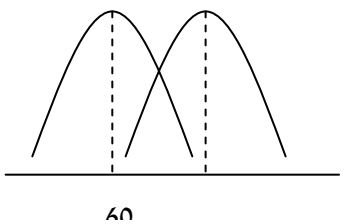
Standard error of estimate

S y.x = S y √ 1 - r 2

זוהי סטית התקן של המנובאים. אם r =1 אזי הניבוי הוא מדויק. אם r = 0 אז אין שום ניבוי. זה כאילו X לא קיים וסטית התקן של הניבוי היא סטית התקן של y.

השערות Hypotheses

קיימת תרופה מסויימת שיעילותה מוכחת. אם תצא תרופה חדשה, יעילותה הטובה יותר היא בחזקת השערה. זוהי אמונה שלא נבדקה עדיין בשטח.



דוגמא נוספת היא ציוני פסיכומטרי בת"א מול אלו של ב"א.ההשערה היא שב"א גבוה יותר מזה של ת"א. יותר מ 600 אך לא ידוע בכמה.

בכדי לקבוע שאכן הציון גבוה יותר אנו נוקטים בשיטה של decision theory . מה הקשר בין הדיווח לאמת. דוגמת האדם שיושב מול מכ"ם ורואה ציפור. האפשרויות שעומדות לפניו הן:

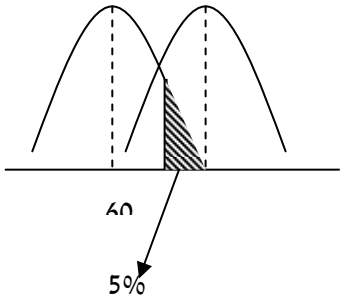
	מטוס	ציפור
ציפור	3	1
מטוס	2	4

אם דיווח ציפור ובאמת זה ציפור - יצא טוב

אם דיווח מטוס וזה באמת מטוס - יצא טוב

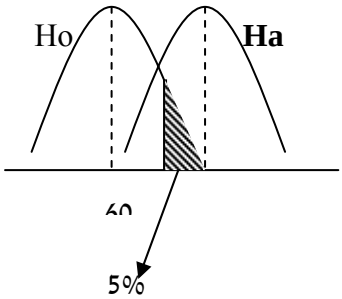
אם דיווח ציפור וזה מטוס - יצא לא טוב

אם דיווח מטוס וזה באמת ציפור - יצא לא טוב.



החוקר את ציוני הפסיכומטרי רוצה להיות בטוח שמה שיצא לו הוא נכון. לכן יש גבול שאם הציון יוצא מעליו הוא בטוח כמעט בודאות. ערך שהוא ערך קריטי. הוא נקבע להיות מעל לשטח 5% של הציון הגבוה של אותם ערכים שצריך להיות מעליהם.

הרצאה 14 - 5/3/1



גרף Ho הוא הגרף הראשון שממנו מתחילים את השוואה לגרף Ha. גרף Ho הוא גרף המיצג את האוכלוסיה, הממוצע שלו הוא ממוצע של אוכלוסיה, $\mu = 550$. גרף Ha הוא גרף המיצג את התפלגות האוכלוסיה המשוער ע"פ מדגם שנערך ע"י החוקר. $\hat{X} = 560$. כדי לקבוע שה"מדגם" גבהו מהאוכלוסיה על ציון הממוצע שלו להיות מעל לערך הקריטי שנקבע.

יתכן שהקביעה ש $\hat{X}$ היא שגויה ולמעשה הוא נמצא בכול זאת מעבר לערך הקריטי בתוך תחום ההתפלגות של Ho, טעות זו נקראת Type one/α error.

יתכן וקבענו ש $\hat{X}$ נמצא מצידו השני של הערך הקריטי לכיוון ממוצע μ וקביעה זו היא שגיאה והערך האמיתי של Ha נמצא מהצד האחד של הערך הקריטי, לכיוון ממוצע $\hat{X}$, טעות זאת נקראת Type two/β error.

היחס בין הטעויות הוא הפוך, ככול שגדל הסיכוי לטעות מסוג α קטן הסיכוי לטעות מסוג β.

הרצאה 15 - 19/3/1

רווח בר סמך Confidence interval

טווח של 95% יתואר בצורה הבאה:

$X^\wedge - 1.96\ S_{x^\wedge} <= \mu <= X^\wedge + 1.96\ S_{x^\wedge}$

למשל נתונים של מחקר מסויים (השפעת אקמול) הם $X^\wedge = 40$ $S = 10$ $n = 25$.

לגבי הקבוצה, לא נשתמש בS של האוכלוסיה אלא בשגיאת התקן של הממוצע $S_{x^\wedge} = 10/5 = 2$. כך שלגבי נתוני האוכלוסיה:

$40 - 1.96 * (2) <= \mu <= 40 + 1.96 * (2)$, הערה - (2) הוא שגיאת התקן.

התוצאה היא $36.08 <= \mu <= 43.92$ כלומר, 95% מהאוכלוסיה יהיו בטווח התוצאות הללו ע"פ תוצאות המדגם.

ככול ש n גדול יותר, המדגם מיצג יותר והטווח של μ יהיה קטן יותר, אנו מתקרבים יותר מ 2 הקצוות אל הממוצע.

Power - הסיכויים/ הסתברות לדחות את Ho למעשה זה 1-β. אם α עולה גם ה power עולה ותחום H1 גדל. תוצאה שיצאה בתוך α נקראת significant result, אם לא היא נקראת not significant result.

T test

$t = X^\wedge - \mu / (S / \sqrt{n})$

סטית התקן כאן מגיעה ממדגם ולא מאוכלוסיה. כול הנתונים מגיעים מהמדגם. הנתון היחיד שהוא לא ממדגם הוא ה μ, זהו נתון קיים.

למשל: מדגם שנתוניו הם $x^\wedge = 520$ $S = 90$ $n = 36$ $\mu = 500$

$520 - 500 / (90 / \sqrt{36}) = 20 / 15 = 1.667$

את הנתון הזה משווים לנתון שבטבלה של t test. בעמודת ה df (דרגות חופש) הנתון הוא n - 1. כך שבמדגם של 36 נבדקים הנתון הוא 35. מאחר ובעמודה שלנו אין 35 נקח את הנתון הנמוך יותר שמופיע והוא 30. בעמודה של 5% הנתון הוא 2.042. היות והנתון שיצא לנו הוא 1.667, הוא נמוך יותר מהערך הקריטי ובמקרה זה נדחה את H_0

אם ה n שואף ל אין סוף, ה t יצא 1.96. כמו בהשערה סטטיסטית של אוכלוסיה מאחר ובמקרה זה המדגם כבר שואף לאוכלוסיה.

הרצאה 16 - 26/3/1

במבחן t סטית התקן היא של המדגם. הגבול כאן הוא ע"פ הטבלה: $t = X^\wedge - \mu / (S / \sqrt{n})$

במבחן z סטית התקן היא של אוכלוסיה והגבולות כאן הם ±1.96:

$Z = X^\wedge - \mu / (\sigma / \sqrt{n})$.

מבחן t של שני מדגמים בלתי תלויים

יש לנו 2 מדגמים ואין נתוני בסיס של אוכלוסיה (למשל 2 תרופות חדשות שיצאו לשוק ולא נבדקו באוכלוסיה עדיין).

$t = X^\wedge_1 - X^\wedge_2 / S_{X^\wedge 1} - S_{X^\wedge 2}$

השערת H_0 במקרה הזה היא $\mu_1 = \mu_2$ או $\mu_1 - \mu_2 = 0$.

השערת H 1 היא $\mu_1 \neq \mu_2$

אנו משתמשים ב μ ולא ב X^ כי הם מיצגים אוכלוסיה.

דרך חישוב המכנה היא :

$S^2 = S_1^2 (N_1 - 1) + S_2^2 (n_2 - 1) / n_1 + n_2 - 2$

נוסחה זו מבטאת את הממוצע של כול המדגמים.

השלב הבא בחישוב במכנה הוא לקחת את ה S^2 שיצא וכדי לקבל את מה שמופיע במכנה לבצע את החישוב הבא :

$\sqrt{S^2 / N_1 + S^2 / N_2}$

שלבים אלו מוצאים S^2 משותף אשר כולל את שני ה S שבמחקר.

את ה t שיצא כאן אני בודק בטבלת מבחן ה t ע"פ דרגות חופש של $(N_1 - 1) + (N_2 - 1)$. את הערך הקריטי שהתקבל אני משווה ל t וממשיך כרגיל עם קבלה או דחיה של H_0.

קיימת דרך נוספת לחשב את המכנה והיא לעשות ישר $\sqrt{S^2 / N_1 + S^2 / N_2}$

אבל אנו לא עושים זאת מאחר ואין כאן חלוקה בשונות משותפת. ההנחה היא שהמדגמים דומים והאוכלוסיות הן שוות (כך גם היא הנחת H_0) ולכן אנו נעשה שונות משותפת. בדוגמה זו השונויות הן נפרדות, שונות אחת מהשניה.

הרצאה 17 - 2/4/01

דוגמא ל t test של 2 מדגמים לא תלויים :

הנתונים : $X^\wedge_1 = 4$ $X^\wedge_2 = 7$ $S^2_1 = 2$ $S^2_2 = 3$ $n_1 = 2$ $n_2 = 36$

הנוסחה היא : $t = X^\wedge_1 - X^\wedge_2 / S\, x^\wedge_1 - x^\wedge_2$ המכנה שווה ל= $\sqrt{S^2/N_1 + S^2/N_2}$

חישוב המונה : 4-7=-3

חישוב המכנה: $S^2 = (n_1 - 1)*S^2_1 + (n_2 - 1)*S^2_2 / N_1 + N_2 - 2 = 48 + 105 / 59 = 2.6$

חישוב כול הנוסחה : $S2 = 4 - 7 / 2.6/25 + 2.6/36 = -3 / 0.4 = -7.5$

יש לבדוק את הערך הקריטי בטבלת t עם דרגת חופשיות מתאימה שהיא $n_1 + n_2 - 2$ כלומר 59.

הנתון המתאים ביותר בטבלה הוא 40. הערך הקריטי הוא 2.021, נתון שהוא ערך מוחלט. ה t שיצא הוא תוצאה מובהקת.

סוגי מבחני השערות סטטיסטיות :

Z test

t test של מדגם אחד.

t test של 2 מדגמים בלתי תלויים

t test מסוג match/ before after

t test מסוג "r"

מחקר לפני אחרי before after או match

מדגם זה הוא לקבוצה שנדגמה בזמן מסויים בעבר ואח"כ נדגמה שוב כדי לבדוק קשר לשינוי מסויים שהתרחש במהלך הזמן שעבר. אפשרות נוספת היא דגימה של 2 קבוצות זהות מכול הבחינות פרט להבדל אחד, מתוך מטרה לבדוק את השפעת הגורם. למשל עישון.

הנוסחה היא : $t = D^\wedge / S_{D^\wedge}$

למשל שיפור מבחני בגרות של נבדקים

4 נבדקים שניגשו פעמיים למבחן בגרות מסויים. B הוא ציון ה "לפני" ו A הוא ציון ה"אחרי"

A	B	D (הפרש B-A)	$D - D^\wedge$
7	8	+1	$(1-(-1))^2 = 4$
10	9	-1	$(-1-(-1))^2 = 0$
8	8	0	$(0-(-1))^2 = 1$
10	6	-4	$(-4-(-1))^2 = 9$
		הממוצע הוא -1	הסה"כ הוא 14

חישוב המונה $D^\wedge$: עמודה D כלומר ממוצע ההפרשים שווה -1.

חישוב המכנה : $S_{D^\wedge} = S_D / \sqrt{n}$.

$\sqrt{n} = 4$, היות ויש לנו 4 נבדקים בטבלה.

$$S_D = \sqrt{\Sigma (D_i - D^\wedge)^2 / \sqrt{N - 1}} = \sqrt{(14 / 3)} = 2.16$$

$$S_{D^\wedge} = S_D / \sqrt{n} = 2.6 / 2 = 1.08$$

חישוב כול הנוסחה : $t = D^\wedge / S_{D^\wedge} = -1 / 1.08 = -0.97$

הערך הקריטי הוא לפי אותה טבלת דרגות חופשיות דו זנבית. חישוב דרגות החופשיות הוא מספר הנבדקים פחות 1, כלומר 3. הערך הקריטי בטבלה הוא 3.182. היות והתוצאה היא -0.97 התוצאה אינה מובהקת ואנו לא דוחים את H_0

הרצאה 18 - 16/4/01

פתרון תרגיל סמסטר ב'.

א. בדיקת ההשערה תעשה ע"י השוואת הממוצעים של 2 הקבוצות

ב. מדובר בהשואה בין 2 מדגמים לא תלויים אחד בשני. בכדי לבדוק האם ההבדל הוא מובהק יש לבצע השערה סטטיסטית :

a. H_0 - ממוצע מספר הפעמים שנשים אומרות תודה שווה לממוצע מספר הפעמים שגברים אומרים תודה.

H_1 - ממוצע מספר הפעמים שנשים אומרות תודה שונה ממוצע מספר הפעמים שגברים אומרים תודה.

b. הגדרת ה α ואיזורי הדחיה וקבלה:

$\alpha = 0.05$ למבחן דו זנבי

הגדרת דרגות החופש $df = 8$

הגדרת t קריטי $t_c = 2.306$

איזור קבלה - $-2.306 <= t_c >= 2.306$

איזור דחיה - $t_c < -2.3306$ וגם $t_c > 2.306$

c. חישוב הסטטיטי

ע"פ הנוסחה $t = X^\wedge_1 - X^\wedge_2 / \sqrt{S^2/N} + \sqrt{S^2/N}$

תחילה יש לחשב את השונות המשותפת המופיעה בנוסחה כ S2. חישובה הוא

ע"פ הנוסחה $S^2 = S^2_1 (N-1) + S^2_2 (N-1) / N_1 + N_2 - 2$

כלומר $34 + 30 / 8 = 8$ זוהי השונות המשותפת.

חישוב ערך ה t $t = 5 - 3 / \sqrt{8/5} + \sqrt{8/5} = 1.12$

d. המסקנה הסטטיסטית היא שהתוצאה אינה מובהקת. $t < 2.306$

ג. עכשיו מדובר בשני מדגמים שכן תלויים אחד בשני. הנוסחה המתאימה היא $t = D^\wedge / SD/\sqrt{n}$. חישוב ה SD בנוסחה זו הוא ע"פ חישוב סטית תקן של הפרשי המדדים. הנוסחה היא $SD = \sqrt{\Sigma (D_i - D^\wedge) / N-1}$.

במקרה זה ההתיחסות היא לא אל כול מדד בפני עצמו אלא אל זוגות של מדדים. לכן דרגות החופשיות כאן הן ע"פ הזוגות. כלומר $df = 4$. כך ש $tc = 2.776$.

$SD = 44/4 = \sqrt{11} = 3.32$

$t = 2 / 3.32/\sqrt{5} = 2 / 1.48 = 1.35$

$t < 2.776$ התוצאה אינה מובהקת.

ד. במקרה זה ההשוואה היא בין מדגם הגברים האמריקאים לבין אוכלוסית הגברים הישראלים. כלומר השוואה בין מדגם לאוכלוסיה. לכן גם נתונים סטית התקן של האוכלוסיה $\sigma = 1$ וממוצע האוכלוסיה $\mu = 2$. יש להשתמש במבחן Z.

הנוסחה היא $Z = X^\wedge - \mu / \sigma/\sqrt{n}$

התוצאה היא $3-2 / 1 / 2.24 = 2.24$

ה Z הקריטי הוא 1.96 כך שתוצאה זו היא מובהקת.

מבחן r

לבדיקה האם מתאם שהתקבל במדגם הוא מובהק או אם לאו.

הנוסחה היא $t = r \sqrt{N - 2 / 1-r^2}$

ההשערות במבחן זה הן

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

למשל במחקר יצאו התוצאות הבאות: $r = 0.6$ ו $n = 60$.

לאחר הצבה פשוטה בנוסחה התוצאה היא 6. את הנתון בודקים בטבלת t לפי דרגות חופשיות של n-2. מאחר ו t_c שווה ל 2 התוצאה במקרה זה היא מובהקת.

הרצאה 19 - 23/4/01

הנחיות לביצוע מבחן t

1. על ההתפלגות להיות נורמלית
2. מספר המבדקים צריך להיות שווה ב 2 הקבוצות $n_1 = n_2$
3. Homoskedascity שויון בין השונויות $s_1 = s_2$. כדי לבדוק זאת ישנו מבחן פשוט מאוד המחלק את השונות הגבוהה יותר בנמוכה יותר. נקרא מבחן F.

מבחן זה הוא גם סוג של השערה סטטיסטית. השערת ה H_0 כאן היא ש F הוא לא מובהק כלומר נמוך מערך שיופיע בטבלה. אם אכן כך ולא דחינו את H_0 אז $s_1 = s_2$ וניתן לערוך את המבחן.

את התוצאה בודקים בתוך טבלה של F כשלכול n נכנסים עם דרגות החופש שלו. כלומר אם ה n של המכנה שווה ל 9, נכנס לטבלה עם המספר 8. התוצאה בטבלה היא של 2 מספרים. המספר העליון יותר מתיחס ל 0.05 והנמוך יותר מתיחס ל 0.01.

מבחן χ^2 - chi square

הנוסחה היא $\chi^2 = \Sigma [(O_i - E_i)^2 / E_i]$

לדוגמא: הטלת מטבע 100 פעמים. ה O מציין את המצוי observed וה E מציין את צפוי/ רצוי expected:

	Oi	Ei
פלי	55	50
עץ	45	50
	100	100

חישוב הנוסחה:

$\chi^2 = \Sigma [(Oi - Ei)^2 / Ei] = (55-50)^2 / 50 + (45-50)^2 / 50 = 1$

הסכומים תמיד יהיו שווים

ההשערות במקרה זה הן:

$H_0 : O_i = E_i$

$H_1 : O_i \neq E_i$

עם הערך 1 נכנסים לטבלה של ערכים קריטים של χ^2 כשאת חישוב ה df עושים בדרך שונה מבדרך כלל. כאן מחשבים את מספר הקטגוריות שיש מינוס אחד. כלומר במקרה של מטבע יש 2 קטגוריות ולכן ה df יהיה שווה ל 1.

התוצאה בטבלה היא 3.84 כלומר לא מובהקת, לא דוחים את H_0. המטבע הוא תקין.

דוגמא נוספת היא של קוביה:

הערך המתקבל מהטבלה הוא 11.07. היות והתוצאה היא 15, הטבלה נמוכה יותר ולכן התוצאה היא מובהקת, דוחים את H_0 והקוביה היא לא מאוזנת.

	Oi	Ei	Oi - Ei
1	10	10	0
2	20	10	$(20-10)^2/10 = 10$
3	5	10	$(5-10)^2/10 = 2.5$
4	5	10	$(5-10)^2/10 = 2.5$
5	10	10	0
6	10	10	0
	60	60	$X^2 = 15$
			df = 5

מבחן מנדל Mendel

מנדל היה נזיר שערך ניסיים בוטנים באפונים. הוא גילה דרך לחשב בצורה דומה למבחן χ^2 ללא הנוסחה לעיל.

הרצאה 20 - 30/4/01

המשך χ^2 -

נקרא גם goodness of fit, טיב ההתאמה בין מה שיצא לבין מה שצריך היה לצאת.

מבחן χ^2 לבדיקת תלות או אי תלות

למשל חלוקה של גברים ונשים במחלקות באוניברסיטה. המבחן יבדוק האם יש קשר בין חלוקת המינים, האם יש תלות ביניהם:

	סה"כ	מחשבים	פסיכולוגיה	כלכלה
גברים	140	50	40	50
נשים	160	30	80	50
	300	80	120	100

הנתונים הללו הם ה Observed. מהו ה expected?

חישוב הצפוי הוא סה"כ הגבר כפול סה"כ כלכלה חלקי כול האוכלוסיה:

140 * 100 / 300 = 47~. נתון צפוי זה של גברים לכלכלה יכנס לטבלה בתא הרלוונטי בפינה הימנית עליונה. את הנתון של נשים בכלכלה ניתן לחשב כמו בטבלאות בייס ע"י חיסור של נתון הגברים מסה"כ נתון הכלכלה. כך שהטבלה תראה כך:

דרגות החופש כאן הן 2 - בכול שורה יש 3 נתונים משתנים. מינוס אחד = 2.

בדוגמה זו $\chi 2 = 18$, ע"פ הטבלה הערך הקריטי הוא 5.99.

ההשערות הסטטיסטיות כאן הן:

$H_0 : O_i = E_i$

	סה"כ	מחשבים	פסיכולוגיה	כלכלה
גברים	140	37 50	56 40	47 50
נשים	160	43 30	64 80	53 50
	300	80	120	100

$H_1 : O_i \neq E_i$

כך שהתוצאה היא מובהקת ויש תלות בין 2 המשתנים.

מאפיינים:

1. במבחן זה הנתונים הם מספריים ולא מדדים כמו בשאר המבחנים האחרים. לכן במבחן זה אין ממוצעים או סטיות תקן. מבחן זה לא מסתמך על התפלגות נורמלית. המספרים כאן הם על רצף, לא חלק מהתפלגות.
2. הסולמות כאן הם של הקטגוריות (מחשבים, פסיכולוגיה וכו') - הם שמיים.

ניתוח שונות Analysis of variance

בדיקת סיבתיות בין נתונים מסויימים, למשל בין ציונים של תלמידים למורה שלהם:

האם יש קשר בין ציונים התלמידים למוריהם?

ציונים					ממוצע
מורה א'	7	9	7	9	8
מורה ב'	10	9	9	10	9.5
מורה ג'	5	6	5	5	5.5

הרצאה 21 - 7/5/01

דוגמא:

בדיקה של רמות כעס בקבוצת נבדקים, ב 4 רמות שונות של טמפרטורה - האם טמפרטורה משפיעה על רמת הכעס. התוצאות הן:

	A1	A2	A3	A4
	1	2	4	6
	1	1	3	6
	2	3	2	5
	4	6	3	3
X^=	2	3	3	5

סימון הטבלה יעשה באופן הבא:

	A1	A2	A3	A4
	X_{11}			X_{41}
	X_{21}			X_{42}
	X_{31}			X_{43}
	X_{41}			X_{44}
$X^{\wedge}_{.}=$	$X_{.1}$	$X_{.2}$	$X_{.3}$	$X_{.4}$
$X^{\wedge}_{..}=$	3.25			

החישוב הוא שונות בין הקבוצות חלקי השונות בתוך הקבוצות. אם השונות "בין" גדולה מהשונות "בתוך" המשמעות היא שיש הבדלים ממשיים בין הקבוצות ולכן יש השפעה של הגורם (המשתנה הבלתי תלוי) על הקבוצה.

	Source	SS	df	MS	F
Between	B		K - 1 = 3		
Within	W		K (n - 1)		
Total	T		(n - 1) = 15		

הדבר הראשון שכדי לחשב במבנה זה הוא את דרגות החופש של ה Total. במקרה של ניסוי הטמפרטורה התוצאה היא 15 = 1 - 16.

נוסחת חישוב ה SS :

$$\Sigma\Sigma(X_{ij} - X^{\wedge}..)^2 = \Sigma\Sigma\ (X_{ij} - X^{\wedge}.j)^2 + \Sigma_{nj}\ (X.j - X^{\wedge}..)^2$$

הרצאה 22 - 14/5/01

המשך נושא קודם

לפי החישוב התוצאות הן : T = 47, B = 19

דרגות החופש של B הן 4 פחות 1 כלומר 3, df של W הם כמות כול הנתונים שיש פחות כמות הקבוצות, כלומר 12 וה df של T הם כמות כול הנתונים שיש פחות 1, כלומר 15.

כך שטבלת הנתונים נראית כך :

Sum square. שונות = Mean square

Source	SS	df	MS	F
B	19	3	19/3=6.33	6.33/2.33=2.71
W	28	12	28/12=2.33	
T	47	15	47/15=3.13	

חישוב הערך הקריטי הוא ע"פ נתוני ה df של W ושל B, כלומר ע"פ 3 במונה ו 12 במכנה. הערך הקריטי הוא 3.49. מאחר והתוצאה היא 2.71 התוצאה היא לא מובהקת ואין קשר. ניתן לנסח זאת או בניסוח P/n s (not significant) או לכתוב P>0.05.

ההשערות ינוסחו כך :

$\mu_1 = \mu_2 = \mu_3 = \mu_4$: H0

$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$: H1

הנוסחה של Sheffe

$$(x_i - x_j)^2 / (S_w^2/n_1) + (S_w^2/n_2)$$

יעשה בה שימוש רק אם התוצאה היא מובהקת. בעזרת נוסחה זו ניתן לאתר בתוך הקבוצות איזה קבוצה היא זו המשפיעה על המובהקות והיטתה את התוצאה לכיוון זה. ניתן להשוות את הקבוצות, אחת לשניה כמה פעמים שרוצים, הדבר לא ישנה את אחוז המובהקות. אם למשל

היתה יוצאת תוצאה מובהקת בדוגמה הנ"ל, החישוב היה נראה כך: השוואה בין קבוצה A1 לבין קבוצה A4 -

$5 - 2 / 2.33/4 + 2.33/4 = 9/1.2 = 7.72$

את הערך הקריטי (3.49) יש להכפיל בדרגות החופש של B, כלומר ב 3. התוצאה היא 10.5, זהו הערך הקריטי המתוקן שאליו משוים את התוצאה (7.72). במקרה זה היא לא מובהקת, כלומר יש סיכוי גדול יותר שהמובהקות נמצאת בין הקבוצות האחרות. כך שכדי להתחיל להשוות בקצוות.

הרצאה 23 - 4/6/01

2 way anova ניתוח שונות דו כיווני

בטבלת הנתונים הבאה יש 2 משתנים בלתי תלויים ומשתנה אחד תלוי. המשתנים הבלתי תלויים הם הקטגוריות. קטגוריה אחת תהיה בטורים, היא תסומן ב A ותחולק (במקרה זה) ל 3 קבוצות (למשל תרופה עם 3 סוגי מינונים שונים). קטגוריה שניה תהיה בשורות, תסומן ב B ובמקרה זה תחולק ל 2 קבוצות (למשל עיתוי לקיחת התרופה). הנתונים שבתוך התאים מיצגים את הנבדקים באותה קבוצה, במקרה זה יש 2 נבדקים בכול תא. אלו הם המשתנים התלויים:

מינוני תרופות

		A1	A2	A3	
עיתוי לקיחת תרופה	B1	2,4	3,5	1,3	$X_{.1.}=3$
	B2	1,1	5,7	7,9	$X_{.2.}=5$
			$X_{..2}=5$		$X_{...}=4$

ע"פ הטבלה ניתן לבדוק:

הבדלים בין שורות (בין זמני התרופות)

הבדלים בין הטורים (הבדלים בין מינוני התרופות)

אינטראקציה - הבדלים בין התאים בטבלה (בין כול קבוצה וקבוצה).

הסימון: Xirc

i - אינדיבידואל, במקרה זה יכול להיות או 1 או 2 מאחר ויש רק 2 בכול תא.

r - שורה, במקרה זה יכול להיות 1 או 2.

c - טור, יכול להיות 1, 2 או 3.

טבלת החישוב:

	source	SS	df	MS	F
between	row	12	1	12	12/1.7
	column	24	2	12	12/1.7
	interaction	32	2	16	16/1.7
	Within	10	6	1.7	

total	78	11

נוסחת החישוב היא:

$$\Sigma\Sigma\Sigma(Xirc - X^{\wedge}...)^2 = nC\Sigma(X^{\wedge}.r. - X^{\wedge}...)^2+nR\Sigma(X^{\wedge}..c - X^{\wedge}...)^2+n\Sigma\Sigma(X^{\wedge}.rc - X^{\wedge}.r. - X^{\wedge}..c + X^{\wedge}...)^2+\Sigma\Sigma\Sigma(Xirc - X^{\wedge}.rc)^2$$

ה n הקטן מסמן את מספר הנבדקים בתא. הסימן X.r. מסמל ממוצע שורה, כך הסימן X..c הוא ממוצע טור, X.rc הוא ממוצע תא ו Xirc הוא כול נתון בכול תא ותא. X... הוא ממוצע הממוצעים.

ההשערות הן 3 קבוצות של השערות לכול שונות ושונות בטבלה:

<u>שונות שורות</u>:

H0: $\mu_{.1.} = \mu_{.2.}$

H1: $\mu_{.1.} \neq \mu_{.2.}$

<u>שונות טורים</u>:

H0: $\mu_{..1} = \mu_{..2} = \mu_{..3}$

H1: $\mu_{..1} \neq \mu_{..2} \neq \mu_{..3}$

<u>שונות אינטראקציה</u>:

H0: $\mu_{.11} = \mu_{.12} = \mu_{.13}$ וכו'...

H1: $\mu_{.11} \neq \mu_{.12} \neq \mu_{.13}$ וכו'...