סיכום קורס סטטיסטיקה ב'
מרצה: עמית נוביק
ע"י שחר דיויס

משפט הגבול המרכזי

נתונות n דגימות, כל דגימה
x_i היא משתנה מקרי כך ש:

$$V(x_i)=\sigma^2 \quad E(x_i)=\mu$$

אם $n\geq30$ מספיק גדול אז

סכום המדגם

$$\boxed{\sum x_i \sim N(n\mu, n\sigma^2)}$$

$$E(\bar{x})=E\left(\frac{1}{n}\sum x_i\right)=\frac{1}{n}E\left(\sum x_i\right)=\frac{1}{n}\cdot n\mu=\mu$$

$$\bar{x}=\frac{1}{n}\cdot\sum_{i=1}^{n}x_i \qquad V(\bar{x})=V\left(\frac{1}{n}\sum x_i\right)=\frac{1}{n^2}V\left(\sum x_i\right)=\frac{1}{n^2}\cdot n\sigma^2=\frac{\sigma^2}{n}$$

ממוצע

$$\boxed{\bar{x}\sim N\left(\mu,\frac{\sigma^2}{n}\right)}$$

מקרה פרטי - בינומי

הסתברות p בהצלחה

$x_i=k$	0	1
$P(x_i=k)$	$1-p$	p
	לא הצלחה	הצלחה

$$x_i=\begin{cases}0 & \text{לא הצלחה}\\ 1 & \text{הצלחה}\end{cases}$$

$$E(x_i)=0\cdot(1-p)+1\cdot p=p$$

$$V(x_i)=E(x_i^2)-[E(x_i)]^2=0^2(1-p)+1^2\cdot p-p^2=p-p^2=p(1-p)$$

אם n גדול מספיק, לפי משפט הגבול המרכזי

$$\sum_{i=1}^{n}x_i\sim N(np, n\cdot p(1-p))$$

נגדיר את x מספר ההצלחות מתוך n דגימות

$$X\sim B(n,p)$$

$$X=\sum x_i$$

במקרה הבינומי

$$\boxed{\begin{array}{l}\text{אם } np>5,\ n(1-p)>5 \\ \text{אז} \\ X\sim N(np, np(1-p))\end{array}}$$

תיקון רציפות

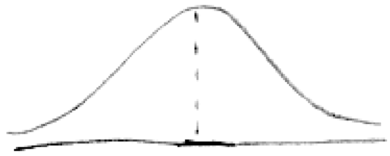
* מה הסיבה?

לקחנו אוכלוסיה עם פרופורציה נתונה p ודגמנו אותה.

$X\sim B(n,p)$ וגם בקירוב: $X\sim N(np, np(1-p))$

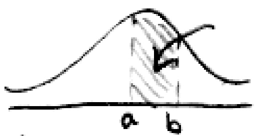
$E(X)=\sum k_i \cdot P(X=k_i)$

$E(X)=np$



$\mu=np$

ההצגה הנורמלית מאפשרת לחשב בקלות את השטח לקבל תחום ערכים



אבל, $X\sim B$ נותנת בדיוק - שטחי עמודים ל- $P(X=k)$

$X\sim N$ נותנת רציף - השטח ל- $P(X=k)$ הוא 0.

ולכן נעשה תיקון רציפות כשנעבור להתפלגות נורמלית:

$P(X=74) = P(73.5<X<74.5)$

ערך בדיד — תיקון רציפות

$P(69\le X<77) = P(68.5<X\le 76.5)$

תיקון רציפות

$\bar{X}=\frac{\sum x_i}{n}=\frac{X}{n}=\frac{\text{מספר הסטודנטים}}{\text{גודל המדגם}}=$ פרופורציית הסטודנטים במדגם $=\hat{p}$

* לא נעשה תיקון רציפות לפרופורציה

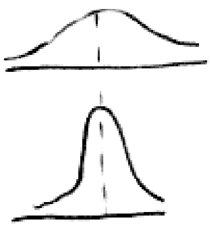
$\bar{X}\sim N(\mu, \frac{\sigma^2}{n})$

אם $np>5$, $n(1-p)>5$ בקירוב

אז

$\hat{p}\sim N(p, \frac{p(1-p)}{n})$

התפלגות $\hat{p}$ / $\bar{X}$ במדגם קטן



במדגמים גדולים

חוק המספרים הגדולים:

ככל שהמדגם גדול יותר

השונות קטנה וצרה יותר

מתכנסים $\hat{p}$ ו- $\bar{X}$

סביב תוחלתם

עד עכשיו הייתה נתונה הפרופורציה באוכלוסיה ונשאלנו
לגבי הסבירות לתוצאות מסויימות של פרופורציות המדגם.
כאשר p, הפרופורציה באוכלוסיה לא ידוע נאמוד אז $\hat{p}$,
הפרופורציה במדגם.

* כש p לא ידוע תנאי למשפט הגבול המרכזי, בגודל המדגם $n \geq 30$

אומד
$\hat{p} = W$
אמד

טעות האמידה $|\hat{p} - p|$

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

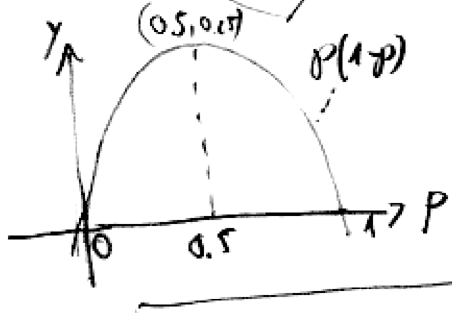
כש p לא ידוע נחליף ב- $\hat{p}$ ואז בקירוב:

$$\underbrace{SE}_{\text{טעות התקן}} = \sigma(\hat{p}) \underset{\text{בקירוב}}{=} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

חוק המספרים הגדולים
ככל שn גדול יותר
סטיית התקן קטנה וסביר
יותר ש $\hat{p}=p$
וגם טעות האמידה
תהיה קטנה יותר.

* כאשר $n \to \infty$
אז $\hat{p} = p$

נמצא את
גבול
התקן
העליון



הסבר לכך ש SE מקסימלי כש $p=\frac{1}{2}$
פרופורציה באוכלוסיה מחולקת חצי-חצי
(בנים, בנות לדוגמא) הסיכוי שבמדגם
תצא פרופורציה שונה גבוה יותר
מאשר אם יש יותר בנים מבנות
או להפך, ככל שיש יותר בנים
באוכלוסיה יש פחות סיכוי לדגום בנות
ולקבל נגיד פרופורציה רחוקה מהפרופורציה
האמיתית.

$$\boxed{SE = \sigma(\hat{p}) \leq \sqrt{\frac{1}{4n}}}$$

---

* אמידת p עם $\hat{p}$: אמידה נקודתית.
* אמידת p עם תחום שבו נצפה למצוא את p בהסתברות בטחון מסויימת:
תחום בר סמך.

רווח סמך של פרופורציה

הפרופורציה המדגמית בסיס לאומדן הפרופורציה באוכלוסיה

$$\hat{p}-\square < p < \hat{p}+\square$$

$$P(\hat{p}-\square < p < \hat{p}+\square) = 1-\alpha$$ רמת הביטחון

נמצא את $\square$:

$$P(p-\square < \hat{p} < p+\square) = 1-\alpha$$

מדגם

$$\hat{p} \sim N(p, SE^2)$$

$$\Phi\left(\frac{p+\square-p}{SE}\right) - \Phi\left(\frac{p-\square-p}{SE}\right) = 1-\alpha$$

$$\Phi\left(\frac{\square}{SE}\right) = 1-\frac{\alpha}{2}$$

$$\frac{\square}{SE} = Z_{1-\frac{\alpha}{2}}$$

רווח סמך עבור p, פרופורציה לא ידועה
על סמך מדגם בגודל n (גדול)
ברמת ביטחון $1-\alpha$:

$$\boxed{\hat{p} - Z_{1-\frac{\alpha}{2}}\underbrace{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}_{SE=\sigma(\hat{p})} \le p \le \hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

מה עושים?

רוצים לדעת מהו p

עושים מדגם

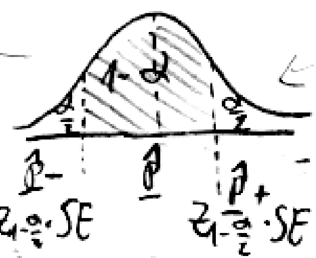
מוצאים $\hat{p}$

מוצאים $\sigma(\hat{p})$ בקירוב (עם $\hat{p}$ במקום p)

↓ חוק המספרים הגדולים

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

התפלגות $\hat{p}$ עם הגבלות:



$$-Z_{1-\frac{\alpha}{2}} = Z_{1-(1-\frac{\alpha}{2})} = Z_{\frac{\alpha}{2}}$$

$\hat{p}$ מתפלג נורמלית סביב התוחלת p
אבל p לא ידוע. אז אנחנו יודעים
ש $1-\alpha$ מערכי $\hat{p}$ נמצאים בתחום
$\hat{p}-Z_{1-\frac{\alpha}{2}}SE$ עד $\hat{p}+Z_{1-\frac{\alpha}{2}}SE$
אז כמו בביטחון של $1-\alpha$
שבתחום זה יימצא p.

* חשוב לציין שהתכונות של
הרווח בר-סמך לא מאוד מדויק
כי σ נגזר מ $\hat{p}$ ש
מאוד ההתפלגות לא מדויק.

אורך רווח הסמך:

$$d = 2Z_{1-\frac{\alpha}{2}} \cdot SE$$

אורך רווח הסמך   רמת הביטחון   סטיית תקן כפונקציה של n

אם רוצים ברמת ביטחון 1-α, רווח סמך באורך d_0 נצטרך n כזאת:

$$2Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{4n}} \le d_0$$

SE מקסימלי

* חצי אורך רווח סמך $\ge |p-\hat{p}|$ טעות הדגימה

$$\hat{p}-\square \qquad \hat{p} \qquad \hat{p}+\square$$

הצגת המקרה הכללי לרווחי סמך של פרמטרים אחרים.

__אם__ א. $O \sim N$ observe - אומד

ב. $E(O) = \theta$ — אומד חסר הטיה - בממוצע ממש שיקבל את ערך הפרמטר

תוחלת האומד (כמו $\hat{p}$)   הפרמטר עצמו (כמו p)

__אז__ רווח סמך ל θ עם האומד O ברמת ביטחון 1-α:

$$\boxed{O - Z_{1-\frac{\alpha}{2}} SE \le \theta \le O + Z_{1-\frac{\alpha}{2}} \cdot SE}$$

רווח סמך עבור תוחלת לא ידועה באוכלוסיה

1) __מדגם גדול__ __סטיית תקן ידועה__

רוצים לאמוד את μ

עושים דגימה ומחשבים את $\bar{X}$, ממוצע המדגם

$n \ge 30$

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

ב. $E(\bar{X}) = \mu$ $\bar{X}$ אומד חסר הטיה ל μ

$$\boxed{\bar{X} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}$$

$$d = 2 \cdot Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

* אורך רווח הסמך
* גבול שגיאת האמידה

$$|\bar{X} - \mu| \leq \frac{1}{2} \cdot d$$

שגיאת האמידה

כמה צריך לדגום כדי שרווח הסמך לא יעלה על אורך מסויים כרצונו בסיכוי $1-\alpha$?

$$2 Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq d_0$$

$$n \geq \left(\frac{2 Z_{1-\frac{\alpha}{2}} \cdot \sigma}{d_0} \right)^2$$

② <u>מדגם קטן</u>
<u>σ ידועה</u>

במדגם קטן לא נוכל להניח נורמליות במשפט הגבול המרכזי. במקרה זה נניח שהאוכלוסיה עצמה מתפלגת נורמלית

✱ <u>משפט</u> אם משתנים מקריים בלתי-תלויים מתפלגים נורמלית גם סכומם והממוצע מתפלג נורמלית.

$X_i \sim N$ הנחה

הנחה בלתי תלוי

$$\sum_{i=1}^{n} x_i \sim N$$

$$\frac{\sum x_i}{n} = \bar{X} \sim N$$

רווח הסמך יהיה כרגיל

③ <u>מדגם גדול</u>
<u>σ לא ידועה</u>

נאמוד את σ מהמדגם:
* במדגמים גדולים S מקבלת ערכים שקרובים ל-σ בסבירות גבוהה. לכן נחליף את σ ב-S בנוסחת רווח הסמך:

$$\boxed{S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - n(\bar{x})^2}{n-1}}$$

תיקון הטיה

$$\boxed{\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}}$$

(4) גודל מדגם קטן, σ לא ידועה

* נניח שהאוכלוסיה מתפלגת נורמלית.

במדגם קטן יש סטיות גדול של S, סטיית התקן המדגמית תהיה רחוקה מ- σ, סטיית התקן האמיתית.
ככל ש n קטן יותר כך יש סטייה גדולה יותר של S מ
רחוקה ל- σ ולכן יש לקחת בחשבון רווח סמך רחב יותר.
לשם כך נשתמש בהתפלגות t, שלוקחת בחשבון את מספר הדגימות: $n-1$ דרגות חופש.

תיקנון

$$\frac{\bar{X}-\mu}{\underbrace{\frac{\sigma}{\sqrt{n}}}_{SE}} = Z \sim N(0,1) \qquad \frac{X-\mu}{\frac{S}{\sqrt{n}}} = t \sim t(n-1)$$

student distribution

* כאשר $n \to \infty$ (שואף) אז $S \to \sigma$ ואז $t \to Z$

רווח סמך עבור μ, תוחלת לא ידועה,
באוכלוסיה בעלת סטיית תקן לא ידועה, עם גודל מדגם קטן מבנה
ברמת בטחון $1-\alpha$:

$$\boxed{\bar{X} - t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2}}(n-1)\frac{S}{\sqrt{n}}}$$

דרגות חופש - הסבר

ככל שהמדגם קטן יותר t גדולה יותר כי S רחוקה יותר מ- σ האמיתית
ולכן המדגם קטן יותר רווח הסמך רחב יותר עבור אותה רמת בטחון.

רווחי סמך להפרשים

הפרש פרופורציות

נדגום שתי אוכלוסיות בלתי-תלויות.

הפרמטר: $\theta = P_x - P_y$

האומד: $O = \underline{\hat{P}}_x - \underline{\hat{P}}_y$

מדגמים ב"ת

מדגם X
$n>30$
$$\downarrow$$
$$\hat{p}_x \sim N\left(p_x, \frac{p_x(1-p_x)}{n}\right)$$

מ. גבול מרכזי

מדגם Y
$m>30$
$$\downarrow$$
$$\hat{p}_y \sim N\left(p_y, \frac{p_y(1-p_y)}{m}\right)$$

אם שתיהם מקרים ב"ת ומדגם
נורמלים אז סכום והפרש משתנים נורמלי

$$\hat{p}_x-\hat{p}_y \sim N(\mu,\sigma^2)$$

$$E(\hat{p}_x-\hat{p}_y)=E(\hat{p}_x)-E(\hat{p}_y)=p_x-p_y$$

$$SE=\sigma(\hat{p}_x-\hat{p}_y)=\sqrt{V(\hat{p}_x-\hat{p}_y)}=\sqrt{V(\hat{p}_x)+V(-\hat{p}_y)+\underbrace{2cov(\hat{p}_x,-\hat{p}_y)}_{cov=0 \text{ ב"ת}}}=\sqrt{V(\hat{p}_x)+V(+\hat{p}_y)}$$

$$\hat{p}_x-\hat{p}_y \sim N\left(p_x-p_y, \frac{p_x(1-p_x)}{n}+\frac{p_y(1-p_y)}{m}\right)$$

רווח סמך להפרש פרופורציות, מדגמים ב"ת, $n,m>30$

$$\boxed{\hat{p}_x-\hat{p}_y \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n}+\frac{\hat{p}_y(1-\hat{p}_y)}{m}}}$$

כי p_x, p_y לא ידועות נחליף בפרופורציות המדגמים

רווח סמך להפרש תוחלות

(1) **מדגמים גדולים ושונות תקן ידועות**

פרמטר: $\theta: \mu_x-\mu_y$

אומד: $\hat{\theta}: \bar{X}-\bar{Y}$

מדגם אוכלוסיה X
$n\geq 30$
$$\downarrow$$
$$\bar{X}$$
$$\downarrow$$
$$\bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$$

משפט גבול מרכזי

מדגם אוכלוסיה Y
$m\geq 30$
$$\downarrow$$
$$\bar{Y}$$
$$\downarrow$$
$$\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{m}\right)$$

$$\downarrow$$

* בנוסף: כשהתנאים הנ"ל חלים שהתפלגות נורמלית
הפרש ממוצעים מתפלג נורמלית.

* גם האוכלוסיות והמדגמים ב"ת זה מזה

$$\bar{X}-\bar{Y} \sim N\left(\mu_x-\mu_y, \sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}\right)$$

② <u>סטיית התקן ידועה, המדגמים קטנים</u>

מניחים שהאוכלוסיה מתפלגת נורמלית

③ <u>סטיית התקן לא ידועה המדגמים גדולים</u>

מחליפים σ ב S. מניחים שהם בסדר קירוב
כי המדגמים גדולים.

④ <u>סטיית התקן לא ידועה, המדגמים קטנים</u>

<u>תנאי</u> $\sigma_x=\sigma_y$ לפי מבחן F משווים של המדגמים
בודקים לפי S

לשתי האוכלוסיות אותה σ נאמוד אותה ע"י שני המדגמים יחד: Spool

$$SE=\sigma(\bar{X}-\bar{Y})=\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}\underset{\sigma_x=\sigma_y}{=}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}=\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}=\boxed{S_p\cdot\sqrt{\frac{1}{m}+\frac{1}{n}}}$$

$$S_p=\sqrt{\frac{\Sigma(x_i-\bar{x})^2+\Sigma(y_i-\bar{y})^2}{n+m-2}}$$

אם כן, ההפרש מתפלג, שתי אוכלוסיות נורמליות
סטיית התקן לא ידועה אך שווה, שני מדגמים ב"ת
ברמת ביטחון $1-\alpha$:

$$\boxed{\bar{X}-\bar{Y}\pm t_{1-\frac{\alpha}{2}}(n+m-2)\cdot S_p\sqrt{\frac{1}{m}+\frac{1}{n}} \leq \mu_x-\mu_y \leq \ldots}$$

* $S_x=\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}} \Rightarrow (n-1)S_x^2=\Sigma(x-\bar{x})^2 \Rightarrow S_p=\sqrt{\frac{(n-1)S_x^2+(m-1)S_y^2}{n+m-2}}$

<u>**מזווגים**</u>

לבדוק האם תרגול משפיע על אוכלוסיה

כשקיים ספק שמא קיים הבדל בין הקבוצות לפני התרגול

דרך 1 בדיקת שתי קבוצות שונות - $\boxed{\text{מדגמים ב"ת}}$
אחת קיבלה טיפול אחר ואחת לא.
פרמטר $\mu_x-\mu_y$ אומד $\bar{x}-\bar{y}$

דרך 2 בדיקת אותה אוכלוסיה לפני ואחרי: $\boxed{\text{מדגמים תלויים}}$

פתרון:

פרמטר μ_D תוחלת הפרש

אומד $\bar{D}_{ifference}$ ממוצע ההפרשים

i	1	2 ...
x_i לפני	100	90 ...
y_i אחרי	100	100 ...
y_i-x_i הפרש	0	10 ...

$$S_D=\sqrt{\frac{\Sigma(D_i-\bar{D})^2}{n-1}}$$

ציון: ממוצע ההפרשים:

משפט: $\bar{D}\sim N(\mu_D, \frac{\sigma_D^2}{n})$

תלוי בגודל המדגם:

מדגם קטן

$$\boxed{\bar{D}-t_{1-\frac{\alpha}{2}}(n-1)\frac{S_D}{\sqrt{n}}\leq\mu_D\leq\ldots}$$

בדיקת השערות

H_0 השערת ה-0 P = הפרופורציה כיום

H_1 ההשערה האלטרנטיבית על הפרופורציה כיום

עושים מדגם ומוצאים את פרופורציית המדגם $\hat{p}$

* אם $\hat{p}$ רחוק מספיק, קטן ← נדחה את H_0
* אם $\hat{p}$ לא רחוק מספיק ← נקבל את H_0

טעות מסוג ראשון

$$P\left(\text{דחיית } H_0 \,\middle|\, H_0 \text{ נכונה}\right) = \alpha$$

ויתור על המוצר טהור
חיפה שלא עברה סף במבחן המציאות

טעות מסוג שני

$$P\left(\text{קבלת } H_0 \,\middle|\, H_0 \text{ שגויה}\right) = \beta$$

$$P\left(\text{נקבל } H_0 \,\middle|\, H_0 \text{ נכונה}\right) = 1-\alpha$$ רמת הביטחון

$$P\left(\text{נדחה } H_0 \,\middle|\, H_0 \text{ שגויה}\right) = 1-\beta$$ עוצמה

החלטה \ מציאות	H_0 נכונה	H שגויה
דחיית H_0	α טעות מסוג ראשון רמת מובהקות	$1-\beta$ עוצמה
קבלת H_0	$1-\alpha$ רמת ביטחון	β טעות מסוג שני
	1	1

2

α ו-β תלויות אחת בשניה:

ככל שנרצה להקטין את טעות מסוג ראשון
נקטין את α (שזו לדחות את H_0 בהנחה שהיא נכונה)
אבל ככל שנתעקש לבדוק ב-H_0 הסיכוי לקבל את H_0
בטעות הוא טעות גדל $= \beta$.

<u>רמת מובהקות</u> אחוז הטעות (מתוך הרבה דגימות) שנרצה
לדחות את H_0 תחת ההנחה ש-H_0 נכונה

<u>הליך במחקר לבדיקת השערות</u>

① נזהה <u>הפרמטר</u> הלא ידוע באוכלוסיה וסימונו

P: שיעור ההצבעה בקרב בני הקיבוצים למועמד הרב"ש במחוז

② ניסוח פורמלי של <u>ההשערות</u>

H_0: $P=0.8$ אין שינוי בשיעור ההצבעה

H_1: $P<0.8$ חשש שמא חלה ירידה בשיעור ההצבעה

③ <u>נתונים</u>: $n=100$

$\hat{p}=0.77$

$\alpha=5\%$ רמת המובהקות

④ נברר <u>התפלגות תחת H_0</u> כדי לקבוע את גבול הסבירות
של $\hat{p}$ ביחס להשערת האפס

בחינת האומד המתאים (ממוצע) שגיל נרמלי:

$np>5$

$n(1-p)>5$

$\downarrow$

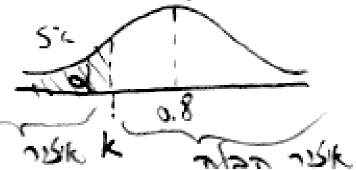
$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$

$\downarrow$ נעבור לציון, התפלגות Z
(כרגיל רק Z)

$$\frac{\hat{p}-0.8}{0.04} = Z \sim N(0,1^2)$$

⑤ שלב <u>ההחלטה</u>: הסקת מסקנות עפ"י הליך הקריטי K

התפלגות $\hat{p}$ תחת H_0:



$$\frac{K-0.8}{0.04} = Z_{0.05}$$

$$\downarrow$$

$$K = 0.734$$

נשווה את הפרופורציה המדגמית לסף הקריטי:

$$K = 0.73 < \hat{p} = 0.77$$

$$\downarrow$$

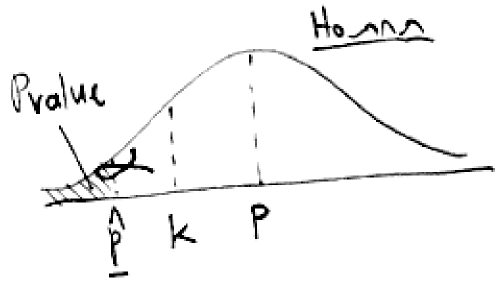
נקבל את H_0

תהליך הבדיקה (ההשערה) קובע האם לקבל או לדחות את H_0
לא האם H_0 נכונה או מוטעית (כי אין עובדה מציאותית)
בחירה ש H_0 נכונה ב 5% מהמקרים נדחה אותה
לב [illegible] נקבלה.

<u>הסקת מסקנה ע"פ רמת מובהקות תוצאת המדגם</u>

Pvalue גודל הקיצוניות של תוצאת המדגם ביחס ל-H_0.

כלומר אזור תוצאה הקיצונית יותר מתוצאת המדגם תחת H_0



אם

$\hat{p} < K$

ז"א

$Pvalue < \alpha$

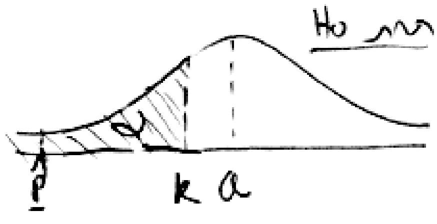
נדחה את H_0

3 מצבים של בדיקת השערות עבור פרמטר θ

(1)

$H_0: \theta = a$

$H_1: \theta < a$



$Pvalue =$ שטח משמאל לתוצאת המדגם

(2)

$H_0: \theta = a$

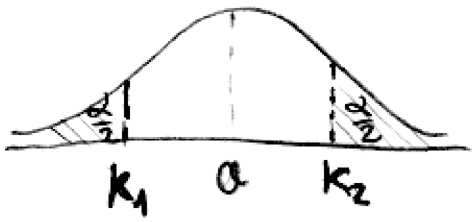
$H_1: \theta > a$

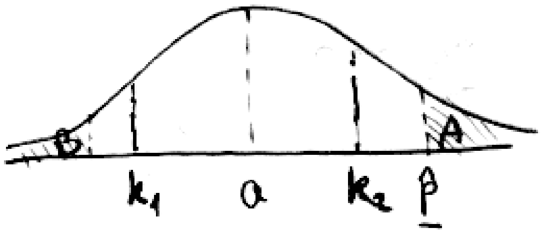
$Pvalue =$ שטח מימין לתוצאת המדגם

(3)

$H_0: \theta = a$

$H_1: \theta \neq a$





$Pvalue = A+B$ איזור התחום שקיצוני יותר מתוצאת המדגם

מציאת הסיכוי לטעות מסוג שני : β

$$\beta = P\left(\text{קבלת } H_0 \;\middle|\; H_0 \text{ שגויה} \right)$$

H_0 נתונה: נמצא את הערך הקריטי שמעליו נדחה את H_0, תחת H_0.
נבדוק את הסיכוי לקבל ערך פחות קיצוני מ- K תחת H_1.

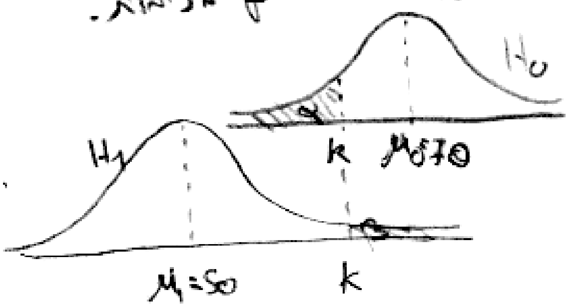
$$\beta = P\left(\text{האומדן פחות קיצוני מ- } K \text{ תחת } H_0 \;\middle|\; H_0 \text{ שגויה, תחת } H_1 \right)$$

* α קובעת את גודל אזור הדחייה:

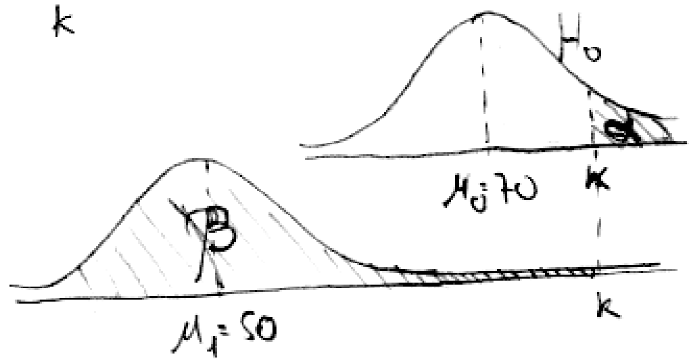
לדוגמא:

* β קובעת את מיקום של אזור הדחייה
נשים שהמיקום החיצוני ביותר תחת H_0 ובכך β מינימלית.

לדוגמא:



ולא:



* כאשר נקבל אזור דחיה וקבלה נבדוק חישוב α + β:

פרמטר: סיכוי ללבן בהטלה בודדת

השערות: חשבנו... H_0: $p=0.5$

פתאום H_1: $p=0.7$

סטטיסטי: הערך "המעניין" בהטלת מטבע עד שמתקבלים.
X- מספר ההטלות עד שהתקבל.

התפלגות הסטטיסטי תחת השערות

H_0: $X \sim G(0.5)$

H_1: $X \sim G(0.7)$

דוגמא שלמה

קביעת אזורי דחיה וקבלה

דחיה $\{X \geq Z\}$ קבלה $\{X < Z\}$ ⇐ חישוב α ו β

הליך מספר למציאת a, b של גרף נתון:

① **פרמטר** — מה שרוצים לגלות: n — מספר המשתתפים בניסוי (המספר שהמכונה לא זוכרת)

② **השערות**

$H_0: n=500$

$H_1: n=400$

③ **סטטיסטי**: X- מספר התלמידים מבין המשתתפים. (הנתון שמסופק במבחן ולפיו נחליט איזו השערה נכונה)

④ **התפלגות הסטטיסטי תחת ההשערות**

$H_0: X \sim B(500, 0.6)$

$np=30$, $n(1-p)=200$ >5

↓ מ. גבול מרכזי

$X \sim N(300, 120)$ — np, $np(1-p)$

$H_1: X \sim B(400, 0.6)$

$np=240$, $n(1-p)=160$ >5

↓ מ. גבול מרכזי

$X \sim N(240, 96)$

⑤ **איזורי דחיה וקבלה**

דחיה: $\{X \leq 299\}$

קבלה: $\{X \geq 300\}$

חסר: מבחני חי בריבוע
לבדיקת אי-תלות וטיב התאמה

דף נוסחאות

רווחי סמך

מה אומדים?	$SE=$ סטיית תקן ידועה — מדגם גדול: מר גבול מרכזי	$SE=$ סטיית תקן ידועה — מדגם קטן: תחת אוכלוסיה נורמלית	$SE\approx$ סטיית תקן לא ידועה (מקורבת SE) — מדגם גדול: גבול מרכזי	$SE\approx$ סטיית תקן לא ידועה — מדגם קטן: תחת אוכלוסיה / שינוי נורמלי
תוחלת μ	$\bar{X}-Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\leq$		$\bar{X}-Z_{1-\frac{\alpha}{2}}\cdot\frac{S}{\sqrt{n}}\leq\dots$	$\bar{X}-t_{1-\frac{\alpha}{2}}(n-1)\cdot\frac{S}{\sqrt{n}}\leq\dots$
הפרש תוחלות $\mu_X-\mu_Y$ מדגמים בלתי תלויים	$(\bar{X}-\bar{Y})-Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\leq$		$(\bar{X}-\bar{Y})-Z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{m}}\leq\dots$	$(\bar{X}-\bar{Y})-t_{1-\frac{\alpha}{2}}(n+m-2)\cdot S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\leq$ סטיית התקן האוכלוסיה שווה ולא ידועה
תוחלת הפרשים μ_D מדגמים תלויים	סטיית התקן לרוב אינה ידועה		$\bar{D}-Z_{1-\frac{\alpha}{2}}\cdot\frac{S_D}{\sqrt{n}}\leq\dots$	$\bar{D}-t_{1-\frac{\alpha}{2}}(n-1)\frac{S_D}{\sqrt{n}}\leq\dots$
פרופורציה p	כשרוצים לבנות רווח סמך לפרופורציה p לא ידועה, ולכן גם SE לא ידועה, נחשב אותה לפי האומד של p		$\hat{p}-Z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\leq\dots$	מאחר ו-p לא ידועה לא ניתן לחשב SE במדגם קטן
הפרש פרופורציות P_X-P_Y מדגמים בלתי תלויים			$(\hat{p}_X-\hat{p}_Y)-Z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}\leq$	

נוסחאות עזר

$$\hat{p}=\frac{X}{n}=\frac{\text{מספר הצלחות}}{\text{גודל המדגם}}\qquad \bar{X}=\frac{\sum_{i=1}^{n}X_i}{n}\qquad S=\sqrt{\frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n-1}}\qquad S_p=\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}$$

בדיקת השערות

1. זיהוי הפרמטר הלא ידוע
2. השערות:

$H_0: P=0.5$ לא שינוי
$H_1: P<0.5$ קטן

שיעור הצלחה במדגם $\hat{p}$: יחסי

3. נתונים: n, $\hat{p}$, α
4. התפלגות האומד תחת H_0:

תנאים לשימוש בגבול המרכזי:

תוחלת: $n<30$ → תחת אוכלוסיה נורמלית; $n\geq30$
פרופורציה: p לא ידועה → $n\geq30$; p ידועה → $np>5$, $n(1-p)>5$

⇓

התפלגות נורמלית של האומד (μ)

$(Z/t,\ S/\sigma)$ → תקן

$\frac{\hat{p}-0.5}{0.04}=Z\sim N(0,1)$

5. החלטה: α / Pvalue
אזורי דחיה וקבלה → החלטה

מציאת α, β במבחן נתון

1. פרמטר
2. השערות
3. סטטיסטי + נתונים
4. התפלגות הסטטיסטי
 תחת H_0
 תחת H_1
5. אזורי דחיה וקבלה
6. חישוב α, β

החלטה \ מציאות	H_0 נכונה	H_0 שגויה	
דחית H_0	α טעות I רמת מובהקות	$1-\beta$ עוצמה	
קבלת H_0	$1-\alpha$ רמת בטחון	β טעות II	
	1	1	2

χ^2 (חי בריבוע) לב התאמה

קטגוריה	O	E

השערות:
תקין / לא תקין
יש קשר / אין קשר
תלוי ...

חישוב:
$$\chi^2=\sum\frac{(O_i-E_i)^2}{E_i}$$

מציאת:
$$\chi^2_{0.95}(n-1)=?$$

החלטה

לוח מוגן: $df=(n-1)(m-1)$

$(n-1)$ לב התאמה
$(n-1)(m-1)$ אי-תלות

ניסויים נוספים בת"ל זה בזה
בכל ניסוי הסיכוי להצלחה זהה - p

בכל ניסוי 2 תוצאות אפשריות בלבד (הצלחה/כשלון)

X מספר הניסיונות עד וכולל הצלחה ראשונה	עורכים n ניסויים → X מספר ההצלחות מבין n נסויים	X מקבל ערכים שלמים בין שני ערכים	$X\sim N(\mu,\sigma^2)$ → $z=\frac{x-\mu}{\sigma}$	
$X\sim G(p)$	$X\sim B(n,p)$	$X\sim U(a,b)$	$Z\sim N(0,1^2)$	
$(1-p)^{k-1}\cdot p$ $P(X>n)=(1-p)^n$ $P(X=m+n \mid X>n)=P(X=m)$	$\binom{n}{k}p^k(1-p)^{n-k}$ $\binom{n}{k}=\frac{n!}{k!(n-k)!}$	$\frac{1}{b-a+1}$	0 $P(X>k)=\Phi(z)$ $\Phi(-z)=1-\Phi(z)$ $-z \quad 0 \quad +z$	$P(X=k)$
$\frac{1}{p}$	$n\cdot p$	$\frac{a+b}{2}$ סימטרית: אמצע התחום	μ	$E(X)=\sum k_i\cdot P(X=k_i)$
$\frac{1-p}{p^2}$	$np(1-p)$		σ^2	$V(X)=E(X^2)-[E(X)]^2=E[X-E(X)]^2$

$P(X=k)$ — הסיכוי שלפרט באוכלוסיה תהיה תכונה מסויימת

$P(X=m)$ הסיכוי שיתקבלו X פרטים בעלי התכונה במדגם נתון

$P(\hat{p}=L)$ הסיכוי שיתקבל פרופורציה מסויימת של התכונה במדגם נתון

$X\sim B(n,p)$

$\sum x_i=X$

$\boxed{np>5 \quad n(1-p)>5}$

נשתמש בגבול המרכזי

$X\sim N(np,\ np(1-p))$

$\hat{p}\sim N\left(p,\ \frac{p(1-p)}{n}\right)$

$\boxed{n\geq 30}$

נשתמש בגבול המרכזי

$\sum x_i\sim N(n\mu,\ n\sigma^2)$

$\bar{x}\sim N\left(\mu,\ \frac{\sigma^2}{n}\right)$

אורך רווח סמך נדרש

$d_0\geq 2Z_{1-\frac{\alpha}{2}}\cdot SE$

במקרה תוחלת

$SE=\frac{\sigma}{\sqrt{n}}$

$n\geq\frac{4Z_{1-\frac{\alpha}{2}}^2\sigma^2}{d_0^2}$

במקרה הבינומי

$SE=\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\xrightarrow{max}$

$d_0\geq 2\cdot Z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{4n}}$

$n\geq\left(\frac{Z_{1-\frac{\alpha}{2}}}{d_0}\right)^2$

מזווגים

i	1	2	...
לפני X_i	100	90	...
אחרי Y_i	100	100	...
ההפרש X_i-Y_i	0	10	...

$\bar{D}=$

$S_D=\sqrt{\frac{\sum(D_i-\bar{D})^2}{n-1}}$