התפלגות המשתנה המקרי

$F(x) \equiv \int_{-\infty}^{x} f(X=x)\,dx \equiv P(X<x)$

תוחלת E(X)

$\mu = \frac{\sum x_i}{N}$

מ"מ בדיד: $E(X)=\sum_{x_i} x_i\cdot P(x_i)=\mu$

מ"מ רציף: $E(X)=\int_{-\infty}^{\infty} x f(x)dx$

$\sigma=\sqrt{\frac{\sum(x_i-\mu)^2}{N}}$

תכונות:

1. $E(a)=a$
2. $E(a+bX)=a+bE(X)$
3. $E(\sum_{i=1}^n X_i)=\sum_{i=1}^n E(X_i)$, $E(X^2-X)=E(X^2)-E(X)$
4. $E(X-\mu)=E(X)-\mu=0$
5. $E(X^2)=V(X)+[E(X)]^2$
6. אם המ"מ בת"ל מתקיים: $E(X_1\cdot X_2)=E(X_1)\cdot E(X_2)$
7. תוחלת של פונ' G

מ"מ בדיד: $E(G(X))=\sum G(x_i)\cdot P(x_i)$

מ"מ רציף: $E(G(X))=\int_{-\infty}^{\infty} G(x)\cdot f(x)dx$

שונות V(X)

$$\boxed{\sigma^2=V(X)=E\left((X-\mu)^2\right)=E(X^2)-\mu^2=E(X^2)-[E(X)]^2}$$

סטיית תקן

$$\boxed{\sigma=\sqrt{V(X)}=\sqrt{\sigma^2}}$$

תכונות של שונות וסטיית תקן:

1. נוסחת העבודה $\mathbf{V(X)=E(X-\mu)^2=E(X^2)-\mu^2}$
2. $\sigma_X\ge 0$, $V(X)\ge 0$
3. שונות/סטיית תקן של גודל קבוע $\sigma_a=0$, $V(a)=0$
4. $V(X+a)=V(X)$ ו- $V(bX)=b^2V(X)$
   $\leftarrow$ $V(a+bX)=b^2V(X)$
5. $\sigma_{a+bX}=|b|\sigma_X$
6. אם המ"מ בלתי תלויים מתקיים: $V(\sum_{i=1}^n X_i)=\sum_{i=1}^n V(X_i)$

התפלגות המקסימום $Y=\mathrm{Max}X_i$

$F_y(Y)=P(Y\le y)=P(MaxX_i\le y)=P(x_1\le y\cap\ldots\cap x_n\le y)$

$\underset{\text{ב"ת}}{=}\prod_{i=1}^n P(x_i\le y)=\prod_{i=1}^n F_{x_i}(y)\underset{\text{ש"ה}}{=}[F_x(y)]^n$

גזירה של $F_y(Y)$ לפי y: $f_y(y)=\frac{F_y(Y)}{dy}=n\cdot[F_x(y)]^{n-1}\cdot f_x(y)$

התפלגות המינימום $Y=\mathrm{Min}X_i$

$F_y(Y)=P(Y\le y)=P(MinX_i\le y)=1-P(MinX_i>y)=$

$=1-P(x_1>y,\ldots,x_n>y)\underset{\text{ב"ת}}{=}1-\prod_{i=1}^n P(x_i>y)=$

$=1-\prod_{i=1}^n\left(1-F_{x_i}(y)\right)\underset{\text{ש"ה}}{=}1-[1-F_x(y)]^n$

גזירה של $F_y(Y)$ לפי y: $f_y(y)=n\cdot[1-F_x(y)]^{n-1}\cdot f_x(y)$

סטטיסטיקה תיאורית – מושגים חשובים

θ פרמטר – מדד מתוך התפלגות האוכלוסייה. גודל קבוע, לא ידוע!

T סטטיסטי – מדד המתקבל מתוך המדגם (פונ' של התצפיות)

$\hat\theta$ – אומד לפרמטר, הינו סטטיסטי. הערך המספרי של $\hat\theta$ נקרא אומדן.

- כשמבקשים למצוא **אומדן** לפרמטר, מוצאים את האומד ומציבים את תוצאות המדגם!

סטטיסטי ממוצע המדגם $\bar x=\frac{1}{n}\sum_{i=1}^{i=n}x_i$

$\sigma_{\bar x}=\frac{\sigma}{\sqrt n}$, $V(\bar x)=\frac{\sigma^2}{n}$, $E(\bar x)=\mu$

(1 **סכום הסטיות** של התצפיות מהממוצע הינו אפס $\sum_{i=1}^{i=n}(x_i-\bar x)=0$

(2 **סכום ריבועי הסטיות** של התצפיות מהממוצע $\sum_{i=1}^{i=n}(x_i-\bar x)^2$ אינו אפס, אך הוא הקטן ביותר מכל מדדי המיקום. (כאשר מציבים את הממוצע זה הערך הכי נמוך שניתן לקבל מסכום ריבועי הסטיות)

התפלגות המדגם המקרי

$f_{X_1,X_2,\ldots,X_n}(X_1,X_2,\ldots,X_n)=\prod_{i=1}^n f_X(X_i)$

משפט הגבול המרכזי

$X_1,\ldots,X_n$ n משתנים מקריים ב"ת וש"ה המקיימים: $E(X_i)=\mu$, $V(X_i)=\sigma^2$

אזי, עבור **n>30**:

סכום המ"מ מתפלג נורמלית כך שמתקיים: $\sum_n X_i\sim N(n\mu,n\sigma^2)$

ממוצעם מתפלג גם הוא נורמלית כך שמתקיים: $\bar X\sim N(\mu,\frac{\sigma^2}{n})$

הערה: אם $X_1,\ldots,X_n$ מתפלגים נורמלית אז **המשפט מתקיים לכל n**.

תיקון רציפות

✓ כאשר משתמשים במשפט הגבול המרכזי עבור אוכלוסייה המקיימת התפלגות **בדידה** (עוברים מ-מ"מ בדיד למ"מ רציף)

✓ עבור $n\le 100$

גם וגם

$P(X\le k)=\varphi\left(\frac{k+0.5-\mu}{\sigma}\right)$

$P(X<k)=P(X\le k-1)=\varphi\left(\frac{k-0.5-\mu}{\sigma}\right)$

$P(X\ge k)=1-P(X<k)=1-\varphi\left(\frac{k-0.5-\mu}{\sigma}\right)$

$P(X>k)=1-P(X\le k)=1-\varphi\left(\frac{k+0.5-\mu}{\sigma}\right)$

$P(X=k)=\varphi\left(\frac{k+0.5-\mu}{\sigma}\right)-\varphi\left(\frac{k-0.5-\mu}{\sigma}\right)$

$P(a\le X\le b)=\varphi\left(\frac{b+0.5-\mu}{\sigma}\right)-\varphi\left(\frac{a-0.5-\mu}{\sigma}\right)$

- לממוצע לא עושים תיקון רציפות, לסכום כן!

קירוב נורמלי להתפלגות הבינומית

יהי X_i מ"מ ברנולי (מ"מ המייצג תוצאה של ניסוי בודד).

הערך 1 מייצג "הצלחה" והערך 0 "כישלון".

X_i	1	0
$P(X_i)$	p	$1-p\equiv q$

$E(X_i)=p$

$V(X_i)=pq$

יהי Y מ"מ בינומי, סכום של n משתני ברנולי

$Y\sim B(n,p)=\sum_{i=1}^{i=n}X_i$

לפי מג"מ:

$Y\sim B(n,p)=\sum_{i=1}^{n}X_i\sim N(np,npq)$

$\bar X=\frac{\sum_{i=1}^{i=n}x_i}{n}=\frac{X}{n}=\hat p\sim N(p,\frac{pq}{n})$

1. $\bar X$ הוא גם פרופורציית (אחוז) ההצלחות שהתקבלו ב-n הניסויים, **ותוחלתו p** ($\sum_{i=1}^{i=n}x_i$ זהו מספר ההצלחות ב-n ניסויי ברנולי)
2. ע"מ להשתמש בקירוב זה, יש לדרוש את התנאי: $np\ge 10$ $nq\ge 10$
3. ההתפלגות הבינומית היא בדידה. לכל n קטן מ-100 נעשה תיקון רציפות.

פונקציית ההתפלגות של המדגם

יהי $X_1,\ldots,X_n$ מדגם מקרי בת"ל וש"ה מתוך אוכלוסייה בעלת התפלגות $f(x)$. **התפלגות המדגם** (=פונקציית ההתפלגות המשותפת) היא:

$\prod_{i=1}^n f(x_i)\,/\,\prod_{i=1}^n P(X=x_i)$

ההסתברות של הריאלזציה, ההסתברות לקבל צירוף מסוים של x_i כתוצאת המדגם.

התפלגות המדגם עבור מדגם מקרי בגודל n שנלקח מהאוכ' הבאות:

א. ברנולי

התפלגות אוכלוסייה $p(x)=p^x(1-p)^{1-x},\ \ \mu=p,\ \ \sigma^2=p*q$

התפלגות המדגם: $\prod p^x(1-p)^{1-x}=p^{\sum x}(1-p)^{n-\sum x}$

ב. אחידה

התפלגות אוכלוסייה: $f(x)=\frac{1}{b-a}$ התפלגות המדגם: $\prod\frac{1}{b-a}=(\frac{1}{b-a})^n$

ג. פואסונית

התפלגות אוכלוסייה: $p(y)=\frac{e^{-\lambda}\lambda^y}{y!}$ התפלגות המדגם: $\prod\frac{e^{-\lambda}\lambda^y}{y!}=\frac{e^{-n\lambda}\lambda^{\sum y}}{\prod y!}$

ד. נורמלית

התפלגות אוכלוסייה: $f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

התפלגות המדגם: $\prod\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=(2\pi\sigma^2)^{-\frac{n}{2}}*e^{-\frac{\sum(x-\mu)^2}{2\sigma^2}}$

ה. גיאומטרית

התפלגות אוכלוסייה: $p(x)=p(1-p)^{x-1}$

התפלגות המדגם: $\prod p(1-p)^{x-1}=p^n(1-p)^{\sum x-n}$

ו. אקספוננציאלית

התפ' אוכלוסייה: $f(x)=\lambda e^{-\lambda x}$ התפ' מדגם: $\prod\lambda e^{-\lambda x}=\lambda^n e^{-\lambda\sum x}$

אמידה נקודתית ($\hat\theta$ - האומד הנקודתי ל-θ)

1. **אמידה בשיטת המומנטים:**

המומנט ה-k באוכלוסייה: $E(X^k)=\mu_k$

המומנט ה-k במדגם: $\frac{\sum_{i=1}^{i=n}x_i{}^k}{n}=m_k$

השיטה מתבססת על השוואת μ_k ל-m_k.

המשוואות המתקבלות עבור המומנט הראשון והשני:

מומנט ראשון: $E(X)=\bar X$

מומנט שני: $V(X)=\frac{\sum_{i=1}^{i=n}x_i{}^2}{n}-(\bar X)^2$, $\boldsymbol{E(x^2)}=\frac{\sum_{i=1}^{i=n}x_i{}^2}{n}$

$$\left[\frac{\sum_{i=1}^{i=n}(x_i-\bar X)^2}{n}=\frac{\sum_{i=1}^{i=n}x_i{}^2}{n}-\bar X^2\ \ ,\ \ E(X^2)=\sigma^2+\mu^2\right]$$

התפלגות אחידה בשיטת המומנטים x~u(a,b)

$\hat b=\bar X+\sqrt{3\left(\frac{\sum_{i=1}^{n}X_i^2}{n}-(\bar X)^2\right)}$ $\hat a=\bar X-\sqrt{3\left(\frac{\sum_{i=1}^{n}X_i^2}{n}-(\bar X)^2\right)}$

2. **<u>אמידה בשיטת הנראות המקסימאלית:</u>**

פונ' הנראות מתארת את ההסתברות לקבל **ריאלציה מסוימת במדגם** כתלות בפרמטר הלא ידוע θ. נשים לב שזו בדיוק <u>התפלגות המדגם</u>.

אנ"מ $\hat{\theta}$ הוא הממקסם את פונ' הנראות: $\hat{\theta} = argmax[L(\theta)]$

שיטת העבודה:

א. נרשום את פונ' הנראות כפונקציה של θ
במקרה הבדיד: $L(\theta) = \prod_{i=1}^{n} P_\theta(X_i = x_i)$
במקרה הרציף: $L(\theta) = \prod_{i=1}^{n} f_\theta(x_i)$

ב. בד"כ נוציא ln מהפונקציה.

ג. נגזור לפי θ ונמצא מקסימום (נציין כי היינו גוזרים פעם שנייה ובודקים שזהו המקסימום). אם יש יותר מאומדן אחד, נגזור את פונ' הנראות לפי כל אומדן.

<u>אנ"מים מוכרים</u>

- <u>אחידה רציפה</u>: כאשר $U(0, \theta)$, אז $\hat{\theta} = max(X_i)$
  כאשר $U(a, b)$, אז $\hat{b} = max(X_i)$, $\hat{a} = \min(X_i)$
- <u>פואסונית</u>: $\hat{\lambda} = \bar{X}$
- <u>בינומית</u>: אם $X \sim B(n,p)$ האנ"מ להסתברות p הוא $\hat{p} = \frac{X}{n}$
- <u>מעריכית</u>: $\hat{\lambda} = \frac{1}{\bar{X}}$
- <u>נורמלית</u>: כש-μ לא ידוע: $\hat{\mu} = \frac{\sum X}{n} = \bar{X}$, $\widehat{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{n}$
  כש-μ ידוע: $\widehat{\sigma^2} = \frac{\sum (X_i - \mu)^2}{n}$

<u>חוקי ln:</u>

$$\ln\left(\frac{x}{y}\right) = lnx - lny \quad , \quad ln(xy) = lnx + lny \quad , \quad \ln(x^k) = klnx$$

$$\frac{d}{dx}\ln(x) = \frac{1}{x} \quad , \quad ln\left(\prod_{i=1}^{n} x_i\right) = \sum_{i=1}^{n} lnx_i$$

<u>אומד חסר הטייה לשונות האוכלוסייה על סמך מדגם בגודל n</u>

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n \cdot \bar{x}^2}{n-1}$$

- על מנת לאמוד את סטיית התקן באוכלוסייה (σ) משתמשים באומד s (מוציאים שורש).
- s אינו חסר-הטייה לסטיית התקן.

<u>משפט: פונקציה של אומד נראות מקסימלית</u>

נתונה פונ' $y = g(\theta)$ ומבקשים למצוא אנ"מ ל-y.
אם $\hat{\theta}$ הוא אנ"מ ל-θ, אז האנ"מ ל-y הינו: $\hat{y} = g(\hat{\theta})$.
לכן, **אם $X \sim B(n,p)$ אזי האנ"מ להסתברות p הוא $\hat{p} = \frac{X}{n}$**

<u>ממוצע ריבועי השגיאה</u> (MSE)

$$MSE = E\left[\left(\hat{\theta} - \theta\right)^2\right] = V\left(\hat{\theta}\right) + \left[E\left(\hat{\theta}\right) - \theta\right]^2$$

- $V\left(\hat{\theta}\right)$ השונות של האומד $\hat{\theta}$
- האיבר $\left(E\left(\hat{\theta}\right) - \theta\right)$ מכונה ההטיה (*Bias*) של האומד $\hat{\theta}$
- **בד"כ נשווה בין אומדים ע"פ ערך ה-MSE שלהם, ברוב המקרים חוסר הטיה משפר MSE.**

<u>תכונות אומדים</u>

I. **<u>חוסר הטיה</u>:** $\hat{\theta}$ יקרא **אומד חסר הטיה** כאשר מתקיים: $E\left(\hat{\theta}\right) = \theta$ ←
$E\left(\hat{\theta}\right) - \theta = 0$ ← ההטיה שווה ל-0 ← $MSE \equiv V\left(\hat{\theta}\right)$
* על מנת שיתקיים חוסר הטיה נדרש להוכיח $E\left(\hat{\theta}\right) = \theta$
* ראינו בכיתה שלכל n ולכל התפלגות מתקיים: $E(\bar{X}) = E(X_i)$
* ולכן, ממוצע המדגם $\bar{X}$ הוא א.ח.ה לתוחלת ההתפלגות $E(\bar{X}) = \mu$
* s^2 א.ח.ה לשונות ההתפלגות $E(s^2) = \sigma^2$
* **טענה**: <u>עבור פונ' ליניארית</u>, אם $\hat{\theta}$ א.ח.ה. ל-θ ו-$G(\theta) = a + b \cdot \theta$, אזי $a + b \cdot \hat{\theta}$ א.ח.ה. ל-$g(\theta)$
* כשמבקשים אח"ה **"על סמך n תצפיות"** זה רמז לכלל **לממוצע**.

II. **<u>עקיבות</u>:** $\hat{\theta}$ ייקרא **אומד עקיב** אם הוא אומד חסר הטיה וכן $\lim_{n \to \infty} V\left(\hat{\theta}\right) = 0$. (תנאי מספיק ולא הכרחי)
* כדי לבדוק עקיבות: 1. נבדוק חוסר הטיה. 2. נבדוק את שונות האומד
<u>חוסר הטיה לעומת עקיבות</u>

אין קשר בין 2 התכונות!

<u>עקיבות</u> מתייחסת לדיוק המדגם כתלות בגודל המדגם, כלומר מה קורה לאומדן כשדוגמים פרטים רבים באותו מדגם. <u>הטייה</u> מתייחסת לתוחלת האומד, כלומר מה קורה לאומדן כשמבצעים מדגמים רבים.

III. **<u>יעילות</u>:** עבור שני אומדים <u>חסרי הטיה</u> $\hat{\theta}_1$ ו-$\hat{\theta}_2$: אם מתקיים $V(\hat{\theta}_1) < V(\hat{\theta}_2)$ נאמר ש-$\hat{\theta}_1$ **יעיל יותר** מ-$\hat{\theta}_2$.
* **אח"ה יעיל יותר** ↔ *MSE* **קטן יותר**

<u>התפלגויות שימושיות</u>

<u>התפלגות נורמלית</u>

אם $X \sim N(\mu, \sigma^2)$, אז $Y = X + const$ מתפלג $Y \sim N(\mu + const, \sigma^2)$
אם $X \sim N(\mu, \sigma^2)$, אז המ"מ $Y = cX$ (קבוע $- c$) מתפלג $Y \sim N(c\mu, c^2\sigma^2)$
אם $X \sim N(\mu, \sigma^2)$, אז המ"מ $Y = \frac{X-\mu}{\sigma}$ מתפלג $Y \sim N(0,1)$
סכום מ"מ נורמליים הוא מ"מ נורמלי עם סכום התוחלות וסכום השונויות.
הפרש מ"מ נורמליים הוא מ"מ נורמלי עם הפרש התוחלות <u>וסכום</u> השונויות.

<u>התפלגות חי-בריבוע χ^2</u> (לא סימטרית)

התפלגות סכום הריבועים של מ"מ נורמליים סטנדרטיים.
אם $i = 1,2,\dots,n$ $X_i \sim N(\mu, \sigma_i^2)$ בלתי תלויים, אז $\sum_{i=1}^{n}\left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n)$
אם $i = 1,2,\dots,n$ $X_i \sim \chi^2(k_i)$ בלתי תלויים, אז $\sum_{i=1}^{n} X_i \sim \chi^2(\sum_i^n k_i)$

<u>התפלגות t</u> (סימטרית)

אם $Z \sim N(0,1)$ ו - $U \sim \chi^2(k)$, ו - Z, U בלתי תלויים אז $\frac{Z}{\sqrt{U/k}} \sim t(k)$

<u>התפלגות של מספר סטטיסטים חשובים</u>

נתונים מ"מ ב"ת ושווי התפלגות $x_i \sim N(\mu, \sigma^2)$ $i = 1, \dots, n$ אזי:

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$$

אם התוחלת לא ידועה:

$$\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{(n-1)}$$

אם σ לא ידועה: $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$ לעומת $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

<u>התפלגות F</u>

- אם $U \sim \chi^2_{(m)}$ ו - $V \sim \chi^2_{(n)}$ ב"ת, אזי $\frac{U/m}{V/n} \sim F(m,n)$
- אם $X \sim F(m,n)$, אזי $\frac{1}{X} \sim F(n,m)$

<u>רב"ס</u>

רב"ס דו צ'	(T_1, T_2)	$P(T_1 < \theta < T_2) = 1 - \alpha$
רב"ס ח"צ תחתון	(T_3, ∞)	$P(T_3 < \theta) = 1 - \alpha$
רב"ס ח"צ עליון	$(-\infty, T_4)$	$P(T_4 > \theta) = 1 - \alpha$

<u>רב"ס עבור התוחלת μ על סמך מדגם של n תצפיות</u>

הנחות ומידע	רב"ס דו צדדי	רב"ס חד צדדי תחתון	רב"ס חד צדדי עליון
σ^2 ידוע. התפלגות נורמלית של האוכלוסייה או n>30	$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}})$	$\mu \in (\bar{X} - Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}, \infty)$	$\mu \in (-\infty, \bar{X} + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}})$
σ^2 אינו ידוע, n>30.	$\mu \in (\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}})$	$\mu \in (\bar{X} - Z_{1-\alpha}\frac{s}{\sqrt{n}}, \infty)$	$\mu \in (-\infty, \bar{X} + Z_{1-\alpha}\frac{s}{\sqrt{n}})$
σ^2 אינו ידוע. n≤30 וכן האוכלוסייה מתפלגת נורמלית.	$\mu \in (\bar{X} - t^{n-1}_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}}, \bar{X} + t^{n-1}_{1-\frac{\alpha}{2}}\frac{s}{\sqrt{n}})$	$\mu \in (\bar{X} - t^{n-1}_{1-\alpha}\frac{s}{\sqrt{n}}, \infty)$	$\mu \in (-\infty, \bar{X} + t^{n-1}_{1-\alpha}\frac{s}{\sqrt{n}})$

<u>אורך רווח הסמך</u> (רלוונטי רק עבור רב"ס דו-צדדי!)

- הרב"ס הינו סימטרי, כאשר במרכזו נמצא ממוצע המדגם.
- המרחק מהממוצע לכל אחד מהגבולות זהה (כי הרב"ס סימטרי).
  מרחק זה מכונה <u>טעות הדגימה</u> וגודלו $\varepsilon = Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
- אורך הרב"ס (מרחק מגבול עליון לתחתון): $L = 2 \cdot \varepsilon = 2 \cdot Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
- גדלים אלו <u>אינם תלויים בתצפיות</u> שהתקבלו בפועל במדגם!
- אורך הרב"ס גדל כאשר:
  1. רמת הסמך $(1 - \alpha)$ גדלה 2. השונות גדלה 3. גודל המדגם קטן

גודל המדגם המינימלי n המבטיח טעות דגימה שאינה עולה על d ברמת סמך $1 - \alpha$:

$$Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le d \Rightarrow \dots \Rightarrow n \ge \left(\frac{Z_{1-\frac{\alpha}{2}} \cdot \sigma}{d}\right)^2$$

<u>התמודדות עם טענות "מישהו טוען ש...." באמצעות רב"סים</u>

- מישהו טוען שהפרמטר **שווה** לערך מסוים C – מקבלים את הטענה אם **רב"ס דו"צ כולל** את הערך.
- מישהו טוען שהפרמטר **נמוך** מערך מסוים C – מקבלים את הטענה אם **רב"ס ח"צ עליון אינו כולל** את הערך (= כל הרב"ס נמוך מ-C).
- מישהו טוען שהפרמטר **גבוה** מערך מסוים C – מקבלים את הטענה אם **רב"ס ח"צ תחתון אינו כולל** את הערך (= כל הרב"ס גבוה מ-C).

בניית רב"ס לפרופורציה p

נתונה אוכלוסייה עם פרופורציה p. במדגם של n תצפיות מהאוכלוסייה, כל תצפית מתפלגת ברנולי עם סיכוי p להצלחה:

$$X_i \sim B(1,p) \;\Rightarrow\; \sum_{i=1}^{n} X_i \equiv X \sim B(n,p)$$

האומדים הנקודתיים ל-p ול-q: $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X}{n}$, $\hat{q} = 1-\hat{p}$

עבור $np \geq 10, nq \geq 10$ מתקיים משפט הגבול המרכזי ואז: $\hat{p} \sim \mathrm{N}\left(p, \frac{pq}{n}\right)$

אז הרב"ס ל-p באמינות $1-\alpha$:

רווח סמך דו צדדי	$p \in \left(\hat{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$
רווח סמך חד צדדי תחתון	$p \in \left(\hat{p} - Z_{1-\alpha}\sqrt{\frac{\hat{p}\hat{q}}{n}}, 1\right)$
רווח סמך חד צדדי עליון	$p \in \left(0, \hat{p} + Z_{1-\alpha}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

גודל המדגם המינימלי n המבטיח טעות דגימה שאינה עולה על d ברמת סמך $1-\alpha$:

$$Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}} \leq d \;\Rightarrow\; n \geq \left(\frac{Z_{1-\frac{\alpha}{2}}}{d}\right)^2 \cdot \hat{p}\hat{q}$$

<u>לצורך מציאת הגודל $\hat{p}\hat{q}$</u>

א. אם קיימת הערכה מוקדמת לפרופורציה האמיתית לפיה $p \leq C$, נבחר את הערכים $\hat{p} = C, \hat{q} = 1 - C$

ב. אם אין הערכה מוקדמת לפרופורציה האמיתית ניקח את הערכים $p = q = 0.5$ (בערכים אלו הפונ' pq מקבלת מקסימום בתחום (0,1)).

- <u>הערה</u>: האומד p נמצא במרכז הרב"ס.

בניית רב"ס לשונות / סטיית התקן

<u>התוחלת ידועה. התפלגות נורמלית של האוכלוסייה:</u>

רב"ס דו צדדי	רב"ס חד צדדי תחתון	רב"ס חד צדדי עליון
$\sigma^2 \in \left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\frac{\alpha}{2},n}}, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\frac{\alpha}{2},n}}\right)$	$\sigma^2 \in \left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{1-\alpha,n}}, \infty\right)$	$\sigma^2 \in \left(-\infty, \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{\alpha,n}}\right)$

<u>התוחלת **איננה** ידועה. התפלגות נורמלית של האוכלוסייה:</u>

רב"ס דו צדדי	רב"ס חד צדדי תחתון	רב"ס חד צדדי עליון
$\sigma^2 \in \left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right)$	$\sigma^2 \in \left(\frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}}, \infty\right)$	$\sigma^2 \in \left(-\infty, \frac{(n-1)s^2}{\chi^2_{\alpha,n-1}}\right)$

1. לעיתים שימושי: $(n-1)s^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2$
2. מכיוון שהתפלגות χ^2 אינה סימטרית, האומד הנקודתי לא יימצא במרכז הרווח. כמו כן, הנוסחאות לגבי אורך הרב"ס אינן רלוונטיות.
3. אם מבקשים רב"ס ל-σ, עושים רב"ס ל-σ^2 ומוציאים שורש.

בדיקת השערות

- **השערת האפס - H_0** (ברירת המחדל, המצב הקיים)
- **השערת האלטרנטיבה - H_1** (המידע החדש המועלה כנגד H_0)

1. <u>אזור הדחייה (C)</u> – תחום תוצאות המדגם שעבורו מחליטים לדחות את H_0 (ולקבל את H_1).
2. <u>אזור הקבלה ($\bar{C}$)</u> – תחום תוצאות המדגם שעבורו מחליטים לא לדחות את H_0.

<u>טעות מסוג ראשון</u> – דחיית H_0 כאשר היא נכונה: $P_{H_0}(C/H_0) \equiv \alpha$

<u>טעות מסוג שני</u> – אי דחיית H_0 למרות ש-H_1 נכונה: $P_{H_1}(\bar{C}/H_1) \equiv \beta$

<u>עוצמת המבחן</u> – ההסתברות של דחייה מוצדקת של H_0, כלומר קבלת H_1 כאשר H_1 אכן נכונה: $P_{H_1}(C/H_1) \equiv 1-\beta$

	H_1 נכונה	**H_0 נכונה**
המבחן אומר לדחות את H_0	אנחנו צודקים!	ביצענו שגיאה מסוג 1
המבחן אומר לקבל את H_0	ביצענו שגיאה מסוג 2	אנחנו צודקים!

* נרצה למזער את α ו-β (כש- α קטנה ← β בהכרח גדל)

* ככל שאזור הדחייה C קטן יותר ← עוצמת המבחן קטנה יותר

* טעות מסוג ראשון חמורה יותר מאשר טעות מסוג שני.

לכן, נקבע <u>רמת מובהקות</u> (חסם עליון ל- (α הנדרשת מהמבחן (נהוג 5%), ותחת חסם זה נחפש מבחן בעל עוצמה מקסימלית (β מינ').

- אם דוחים את H_0 ברמת מובהקות α, אז נדחה את H_0 גם עבור כל רמת מובהקות $\alpha < \tilde{\alpha}$.
- אם לא דוחים את H_0 ברמת מובהקות α, אז לא נדחה את H_0 גם עבור כל רמת מובהקות $\alpha > \tilde{\alpha}$.

מבחן חד-צדדי

- מבחן חד-צדדי ימני: $H_0: \mu = \mu_0$, $H_1: \mu > \mu_0$ או $H_0: \mu \leq \mu_0$, $H_1: \mu > \mu_0$
- מבחן חד-צדדי שמאלי: אי השוויון של H_1 הפוך.

בדיקת השערות במודל נורמלי

<u>בדיקת השערות על תוחלת</u>

מודל נורמלי – מתקיים כאשר סטטיסטי המבחן (ממוצע המדגם) מתפלג נורמלית, כלומר כאשר מתקיים לפחות אחד מהשניים:

1. האוכלוסייה מתפלגת נורמלית
2. מתקיימים התנאים לשימוש במג"מ (n>30 או nq,np>10)

- כשנתון משקל בד"כ נניח לגביו התפלגות נורמלית

	מבחן חד צדדי		מבחן דו צדדי
מערכת ההשערות	$H_0: \mu \leq \mu_0$ $H_1: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ $H_1: \mu \neq \mu_0$
אזור דחייה	$\bar{x} > \underbrace{\mu_0 + Z_{1-\alpha}\cdot\sigma/\sqrt{n}}_{\text{הערך הקריטי k}}$	$\bar{x} < \mu_0 - Z_{1-\alpha}\cdot\sigma/\sqrt{n}$	$\bar{x} > \mu_0 + z_{1-\frac{\alpha}{2}}\cdot\sigma/\sqrt{n}$ **או:** $\bar{x} < \mu_0 - z_{1-\frac{\alpha}{2}}\cdot\sigma/\sqrt{n}$
גודל מדגם מינ' המבטיח α ו- $1-\beta$ רצויים	$n \geq \left[\frac{(z_{1-\alpha}+z_{1-\beta})\cdot\sigma}{\mu_0-\mu_1}\right]^2$		$n \geq \left[\frac{(z_{1-\frac{\alpha}{2}}+z_{1-\beta})\cdot\sigma}{\mu_0-\mu_1}\right]^2$
לאחר שבוצע מדגם והתקבלה תוצאה מדגמית $\bar{X} = C$ אזי עבור תוצאה זו:			
P-value (מובהקות התוצאה)	$1-\phi\left(\frac{C-\mu_0}{\sigma/\sqrt{n}}\right)$	$\phi\left(\frac{C-\mu_0}{\sigma/\sqrt{n}}\right)$	$2\cdot\left(1-\phi\left(\frac{\lvert C-\mu_0\rvert}{\sigma/\sqrt{n}}\right)\right)$

☒ הטבלה היא עבור שונות ידועה. אם **השונות לא ידועה** אומדים אותה באמצעות s^2 ומחליפים את ההתפלגות הנורמלית בהתפלגות $t^{(n-1)}$.

<u>מובהקות התוצאה P-value</u>

ההסתברות לקבל תוצאה כמו שהתקבלה במדגם בפועל או תוצאה קיצונית ממנה (בכיוון האלטרנטיבה), <u>בהנחה שהשערת האפס נכונה</u>.

* אם התוצאה שהתקבלה במדגם בפועל (C) היא הערך הקריטי של המבחן (K) אז P-value היא בדיוק ההסתברות לטעות מסוג ראשון!

* כלל ההחלטה: אם ערך PV קטן מ- α בו אנו מעוניינים (נאמר שתוצאת הניסוי היא "<u>מובהקת</u>") נדחה את H_0, אחרת לא נדחה את H_0.

P-value היא **רמת המובהקות המינימאלית** שעבורה יוחלט לדחות את H_0.

* אם על סמך מדגם ניתן לדחות את H_0 עבור רמת מובהקות α, H_0 תידחה גם עבור כל רמת מובהקות $\alpha < \tilde{\alpha}$ (כי $PV \leq \alpha < \tilde{\alpha}$).

* אם על סמך מדגם ניתן לקבל את H_0 עבור רמת מובהקות α, H_0 תתקבל (לא תידחה) גם עבור כל רמת מובהקות $\alpha > \tilde{\alpha}$ (כי $PV \geq \alpha > \tilde{\alpha}$).

<u>בדיקת השערות על פרופורציה</u>

נדרש $np_0, nq_0 > 10$ ע"מ שיתקיים מג"מ, ואז מקבלים: $\hat{p} = \frac{X}{n} \sim N\left(p, \frac{pq}{n}\right)$

	מבחן חד צדדי		מבחן דו צדדי
מערכת ההשערות	$H_0: p = p_0$ $H_1: p > p_0$	$H_0: p = p_0$ $H_1: p < p_0$	$H_0: p = p_0$ $H_1: p \neq p_0$
אזור דחייה	$\hat{p} > p_0 + z_{1-\alpha}\cdot\sqrt{\frac{p_0q_0}{n}}$	$\hat{p} < p_0 - z_{1-\alpha}\cdot\sqrt{\frac{p_0q_0}{n}}$	$\hat{p} > p_0 + z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{p_0q_0}{n}}$ **או:** $\hat{p} < p_0 - z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{p_0q_0}{n}}$
גודל מדגם מינ' המבטיח α ו- $1-\beta$ רצויים	$n \geq \left[\frac{Z_{1-\alpha}\sqrt{p_0q_0}+Z_{1-\beta}\sqrt{p_1q_1}}{p_0-p_1}\right]^2$		$n \geq \left[\frac{Z_{1-\frac{\alpha}{2}}\sqrt{p_0q_0}+Z_{1-\beta}\sqrt{p_1q_1}}{p_0-p_1}\right]^2$
לאחר שבוצע מדגם והתקבלה תוצאה מדגמית $\bar{X} = C$ אזי עבור תוצאה זו:			
P-value (מובהקות התוצאה)	$1-\phi\left(\frac{C-p_0}{\sqrt{p_0q_0/n}}\right)$	$\phi\left(\frac{C-p_0}{\sqrt{p_0q_0/n}}\right)$	$2\cdot\left(1-\phi\left(\frac{\lvert C-p_0\rvert}{\sqrt{p_0q_0/n}}\right)\right)$

<u>בדיקת השערות על שונות</u> הנוסחאות בדף המבחן (II)

* מתקיים רק כשהאוכלוסייה מתפלגת נורמלית

* **רמת המובהקות הגבולית**: להתפלגות χ^2 אין נוסחה לחישוב *P-value*. נסתכל בטבלת ההתפלגות במספר דרגות החופש המתאימות ונראה עבור איזה ערך P (1-α) היינו דוחים את ההשערה. ר"ה הגבולית היא בין ערך ה- α הזה לזה שבא אחריו.

<u>בדיקת השערות על שיוויון שונויות באוכלוסיות נורמליות</u> בדף (VI)

1. $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}x_i^2 - n\cdot\bar{x}^2}{n-1}$
2. בטבלת F נתונים לנו רק הערכים עבור $f_{0.95}$, כלומר עבור $1-\frac{\alpha}{2} = 0.95$. לכן, **<u>נשתמש תמיד במובהקות</u> $\alpha = 10\%$**.
3. ע"מ למצוא את $f_{\frac{\alpha}{2}}$ ניעזר בזהות: $f_{\frac{\alpha}{2}}^{\,n_1-1,n_2-1} = \frac{1}{f_{1-\frac{\alpha}{2}}^{\,n_2-1,n_1-1}}$
4. אם לא נדחה את H_0 בר"מ 10%, גם לא נדחה אותה בר"מ קטנה יותר.

<u>בדיקת השערות על הפרש פרופורציות</u> הנוסחאות בדף המבחן (V)

המרה לח"צ ימני: $H_0: p_1 = p_2$

(שמאלי: הופכים סימן) $H_1: p_1 > p_2$

דחה אם: $Z_{\hat{p}_1-\hat{p}_2} > Z_{1-\alpha}$

בדיקת השערות על הפרש תוחלות

הנוסחאות בדף המבחן (IV)

מערכת ההשערות: $H_0: \mu_1 - \mu_2 = d_0$
$H_1: \mu_1 - \mu_2 \neq d_0$

- נתונה הערכה לגבי גודל ההפרש = d_0, **אחרת** $d_0 = 0$

1. <u>מערכת ההשערות</u> – בדף הכלל הוא עבור דו"צ. **<u>עבור מבחן חד צדדי</u>**:
   - מבחן חד"צ שמאלי, דחה אם: $T_D < -t_{1-\alpha}^{(n_1+n_2-1)}$ - $H_1: \mu_1 - \mu_2 < 0$
   - מבחן חד"צ ימני, דחה אם: $T_D > t_{1-\alpha}^{(n_1+n_2-1)}$ - $H_1: \mu_1 - \mu_2 > 0$
2. <u>תלות</u> – האם המדגמים בלתי תלויים או תלויים (=מזווגים). **מדגמים הם תלויים** כאשר זהות הפריטים שנבחרו במדגם אחד מושפעת מזהות הפריטים שנבחרו במדגם השני. המקרים הנפוצים:
   א. כאשר **בשני המדגמים** מופיעים **אותם** נבדקים (אנשים, מכונות...)
   ב. כשיש **התאמה מכוונת** בין הנבדקים במדגמים (למשל: בגיל ומין).
3. <u>שיווין שונויות</u> – אם השונויות בשתי האוכלוסיות אינן ידועות, האם ניתן להניח שהן שוות? <u>איך מחליטים</u>: (1) זה נתון; (2) מבצעים בדיקת השערות על שוויון שונויות; (3) בודקים באמצעות רב"ס ליחס השונויות.

<u>רב"ס להפרש תוחלות ברמת סמך $1-\alpha$</u>

הנחות ומידע	רב"ס דו צדדי	
מדגמים ב"ת, שונויות ידועות $E(X_i) = \mu_1, V(X_i) = \sigma_1^2$ $E(Y_j) = \mu_2, V(Y_j) = \sigma_2^2$ $n_1, n_2 > 30$ או נורמליות	$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\ ,\ \bar{x} - \bar{y} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$	
מדגמים ב"ת, שונויות לא ידועות אך שוות בשתי האוכלוסיות. $X_1, \ldots X_{n_1} \sim N(\mu_1, \sigma^2)$ $Y_1, \ldots Y_{n_2} \sim N(\mu_2, \sigma^2)$	$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - t_{1-\frac{\alpha}{2}}^{n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\ ,\ \bar{x} - \bar{y} + t_{1-\frac{\alpha}{2}}^{n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$	כאשר: $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$
מדגמים ב"ת, שונויות לא ידועות ולא שוות. $X_1, \ldots X_{n_1} \sim N(\mu_1, \sigma_1^2)$ $Y_1, \ldots Y_{n_2} \sim N(\mu_2, \sigma_2^2)$	$\mu_1 - \mu_2 \in \left(\bar{x} - \bar{y} - t_{1-\frac{\alpha}{2}}^{v} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\ ,\ \bar{x} - \bar{y} + t_{1-\frac{\alpha}{2}}^{v} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right)$	כאשר: $v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

<u>רב"ס להפרש פרופורציות</u>

נסתפק בתנאים $n_1q_1 \geq 10, n_1p_1 \geq 10$, וגם $n_2q_2 \geq 10, n_2p_2 \geq 10$:

$$p_1 - p_2 \in \left(\hat{p}_1 - \hat{p}_2 - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\ ,\ \hat{p}_1 - \hat{p}_2 + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right)$$

* הרב"סים סימטריים, וכל מה שלמדנו לגבי אורך רב"סים תקף גם כאן.

<u>בניית רב"ס ליחס שונויות</u>

דרך חלופית לבדיקת טענות שוויון אוכלוסיות: אם השונויות שוות היחס 1. אם הערך 1 נמצא בתוך הרב"ס, נקבע שהשונויות שוות ברמת מובהקות α.

$$\frac{s_1^2 / s_2^2}{f_{1-\frac{\alpha}{2}}^{n_1-1, n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2 / s_2^2}{f_{\frac{\alpha}{2}}^{n_1-1, n_2-1}}$$

<u>מבחן חי בריבוע לטיב התאמה</u>

האם לפי נתוני המדגם סביר שהאוכלוסייה מתפלגת בהתפלגות מסוימת.
השערת האפס היא כי האוכלוסייה מתפלגת בהתפלגות תיאורטית כלשהי.
השערת האלטרנטיבה היא כי התפלגות האוכלוסייה היא אחרת.
* במקרה זה האינפורמציה החדשה "תאושר" ע"י <u>קבלת</u> H_0!

<u>שלבי העבודה</u>

1. מוודאים שהפרמטרים של ההתפלגות ידועים. אם לא, יש לאמוד אותם.
2. יש לנסח בבירור את ההשערות הנבדקות
3. מחלקים את נתוני המדגם לקבוצות (מספר הקבוצות מסומן ב-$\boldsymbol{k}$):
   - אם נתונה חלוקה לקבוצות, מוודאים שתוחלת מס' הפריטים **הצפויים** בכל קבוצה i מקיים: $E_i \geq 5$. אם לא, מאחדים עם קבוצה סמוכה.
   $$(\text{בדיד})\ E_i = n\left[\sum_{x=a}^{b} P(X = x)\right], (\text{רציף})\ E_i = n[P(a \leq X \leq b)]$$
   - אם לא נתונה חלוקה, ניצור אותה בעצמנו כך שהקבוצות יכסו את כל התחום של ההתפלגות, ויתקיים כמו קודם: $E_i \geq 5$.
4. O_i - מס' הפריטים בכל קבוצה כפי שהתקבלו **בפועל במדגם**.
5. חישוב סטטיסטי המבחן: $\chi_{emp}^2 = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$
   עבור מדגם גדול, וכאשר מתקיים $E_i \geq 5$ הסטטיסטי מתפלג בקירוב χ^2 עם ***(k-p-1)*** דרגות חופש, כש-***p*** הוא מס' הפרמטרים שנאמדו. אם לא נאמדו כלל פרמטרים, *p=0*.
6. הגדרת המובהקות הרצויה α ומציאת <u>הערך הטבלאי:</u> $\chi_{1-\alpha,(k-p-1)}^2$
   משמעות הערך הטבלאי: בהסתברות של 1-α נקבל את הערך הטבלאי או ערך קטן ממנו, כאשר השערת האפס מתקיימת.
7. כלל החלטה: דחה את H_0 אם $\chi_{emp}^2 > \chi_{1-\alpha,k-p-1}^2$

<u>מבחן חי בריבוע לאי תלות</u>

בודק האם קיימת תלות בין שני משתנים (בין שתי אוכלוסיות, בין שתי תכונות באותה אוכלוסיה וכו').
נתונות n תצפיות בלתי תלויות, בכל תצפית נרשמו זוג ערכים (X ו-Y).
מסווגים את התצפיות בלוח שכיחויות דו ממדי בעל r שורות ו- c עמודות.
O_{ij} - שכיחות התצפית המוגדרת ע"פ התא (i,j).
בשולי הלוח – השכיחויות של הקטגוריות של כל משתנה בנפרד.

X, Y	Y_1	Y_2	...	Y_j	...	Y_c	סה"כ
X_1	O_{11}	O_{12}	...	O_{1j}	...	O_{1c}	$f_{i\cdot} = \sum_{j=1}^{j=c} O_{ij}$
X_2	O_{21}	O_{22}	...	...	...	...	$f_{2\cdot}$
...	...	...	...	...	...	...	...
X_i	O_{i1}	...	...	O_{ij}	...	O_{ic}	...
...	...	...	...	...	...	...	...
X_r	O_{r1}	...	...	O_{rj}	...	O_{rc}	$f_{r\cdot}$
סה"כ	$f_{\cdot 1} = \sum_{i=1}^{i=r} O_{i1}$	$f_{\cdot 2}$	...	...	...	$f_{\cdot c}$	n

H_0: $P(X_i \cap Y_j) = P(X_i) \cdot P(Y_j) \Leftrightarrow$ לא קיימת תלות בין משתנה השורות למשתנה העמודות

H_1: אחרת

סטטיסטי המבחן לאי תלות: $\chi_{emp}^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

כאשר: $E_{ij} = \frac{f_{i\cdot} \times f_{\cdot j}}{n}$

ע"מ להשתמש במבחן נדרש שבכל תא (i,j) יתקיים $\boldsymbol{E_{ij} \geq 5}$.

כלל ההחלטה ברמת מובהקות α: דחה אם $\chi_{emp}^2 > \chi_{1-\alpha,[(r-1)\cdot(c-1)]}^2$

<u>רגרסיה ליניארית פשוטה</u>

מאפשרת לחזות את ערכו של משתנה אחד ע"פ ערכו של משתנים אחרים.
Y – המשתנה התלוי / המנובא / המוסבר
X – המשתנה הבלתי תלוי / המנבא / המסביר

<u>מודל הרגרסיה הליניארית הפשוטה</u>

הערך הממוצע של *Y* (<u>התוחלת</u>): $E(Y/X = x_i) = \beta_0 + \beta_1 x_i$
עבור פרט i מסוים באוכלוסייה מתקיים: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, כאשר ε_i מייצג "רעש" אקראי קטן. מניחים שהרעשים ב"ת ונורמליים - $\varepsilon_i \sim N(0, \sigma^2)$
לכל i, ולכן מתקיים: $(y_i | X = x) \sim N(\beta_0 + \beta_1 x, \sigma^2)$

<u>מודל הרגרסיה הליניארית במדגם</u>

אומד ל-β_0 - b_0 (החותך) ; אומד ל-β_1 – b_1 (השיפוע). **השגיאה** ε_i לא ידועה.
קו הרגרסיה שיתקבל **מהמדגם**: $\boxed{\hat{y}_i = b_0 + b_1 x_i}$

<u>מקדמי הרגרסיה – שיטת הריבועים הפחותים</u>

הסטייה של תצפית במדגם ביחס לקו נתון $\hat{y}_i$: $e_i = y_i - \hat{y}_i$

$$\boxed{b_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \equiv \frac{SS_{xy}}{SS_x} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\cdot\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} \quad \boxed{b_0 = \bar{y} - b_1\bar{x}}$$

<u>תכונות קו הריבועים הפחותים</u>

- הקו עובר דרך נקודת הממוצעים $(\bar{x}, \bar{y})$
- **האומדים *b₁,b₀* הם חסרי הטיה. כלומר:** $\boldsymbol{E(b_1) = \beta_1 \quad E(b_0) = \beta_0}$

<u>משפט פירוק השונויות</u> - $\boxed{SST = SSE + SSR}$

$$SST \equiv SS_y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$
$$SSR \equiv \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (\text{ברל"פ})\ b_1^2 SS_x$$
$$SSE \equiv \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- ככל שקו הרגרסיה מתאים יותר למודל, SSR יותר גדול ו-SSE יותר קטן.
- בהתאמה מושלמת, $SSR = SST, SSE = 0$
- **"אחוז השונות המוסברת", "מקדם ההסבר":** $R^2 \equiv \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- **המתאם בין שני משתנים מקריים במדגם:** $\hat{\rho} \equiv r = \frac{SS_{xy}}{\sqrt{SS_x \cdot SS_y}}$
- **ברגרסיה ליניארית פשוטה:** $r^2 = R^2$

<u>אמידת σ^2 (השונות באוכלוסיה) ושונויות האומדים b_1, b_0</u>

א.ח.ה ל- σ^2: $\boldsymbol{s^2 = \hat{\sigma}^2 = MSE = \frac{SSE}{n-2} = \frac{SST - SSR}{n-2} = \frac{SS_y - b_1^2 SS_x}{n-2}}$

א.ח.ה לסטיית התקן של *b₁,b₀*: $S_{b_1} = \frac{S}{\sqrt{SS_x}}\ ,\ S_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$

<u>בניית רווח סמך לפרמטרים שנאמדו</u>

רב"ס דו-צדדי ל-β_1 ברמת סמך $1-\alpha$: $\beta_1 \in \left(b_1 \pm \frac{S}{\sqrt{SS_x}} t_{1-\frac{\alpha}{2}}^{(n-2)}\right)$

- אם <u>אפס</u> נמצא ברב"ס הזה, לא ניתן להגיד ברמת סמך $1-\alpha$ כי <u>**הרגרסיה מובהקת** = הקשר בין x ל-y הוא לינארי</u>.

רב"ס דו-צדדי ל-β_0 ברמת סמך $1-\alpha$: $\beta_0 \in \left(b_0 \pm S\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}} t_{1-\frac{\alpha}{2}}^{(n-2)}\right)$

רב"סים ח"צ: נחליף את $1-\frac{\alpha}{2}$ ל- $1-\alpha$ ונעדכן את הגבול המתאים ל- $\pm\infty$.

בדיקת השערות על ערכי β_0, β_1 ברמת מובהקות α

	מבחן חד צדדי		מבחן דו צדדי
מערכת ההשערות	$H_0: \beta_0 = \mu_0$ $H_1: \beta_0 > \mu_0$	$H_0: \beta_0 = \mu_0$ $H_1: \beta_0 < \mu_0$	$H_0: \beta_0 = \mu_0$ $H_1: \beta_0 \neq \mu_0$
אזור דחייה	$T_{b_0} = \frac{b_0 - \mu_0}{S_{b_0}} > t_{1-\alpha}^{n-2}$	$T_{b_0} = \frac{b_0 - \mu_0}{S_{b_0}} < -t_{1-\alpha}^{n-2}$	$\lvert T_{b_0} \rvert = \left\lvert \frac{b_0 - \mu_0}{S_{b_0}} \right\rvert > t_{1-\frac{\alpha}{2}}^{n-2}$

עבור β_1 נציב בהתאם μ_1, b_1 ו- S_{b_1}.

בדיקת השערות על מובהקות הרגרסיה

מערכת ההשערות: $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$ (אם נקבל את H_0 הרגרסיה לא מובהקת)

דרך ראשונה: מבחן t על הפרמטר β_1

סטטיסטי המבחן: $T_{b_1} = \frac{b_1 - 0}{S_{b_1}} = \frac{b_1}{s/\sqrt{SS_x}} \sim t(n-2)$

כלל ההחלטה עבור רמת מובהקות α: דחה את H_0 אם $\lvert T_{b_1} \rvert > t_{1-\frac{\alpha}{2}}^{(n-2)}$

דרך שנייה: מבחן *F*

סטטיסטי המבחן: $F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/n-2}$

כלל ההחלטה עבור רמת מובהקות α: דחה את H_0 אם $F > f_{1-\alpha}^{1,(n-2)}$

- עבור רגרסיה פשוטה בלבד מתקיים הקשר: $F = (T_{b_1})^2$

חיזוי ערכי המשתנה המוסבר

- **רב"ס לאומדן y_p בהינתן $X = x_p$ ברמת סמך $1 - \alpha$:**

$$y_p \in \left((b_0 + b_1 \cdot x_p) \pm t_{1-\frac{\alpha}{2}}^{n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_x}} \right)$$

- **רב"ס לתוחלת (המותנית) $E(y_p)$ בהינתן $X = x_p$ ברמת סמך $1 - \alpha$:**
אותו דבר, רק בלי 1 בתוך השורש

רגרסיה לינארית מרובה $\boxed{y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i}$

- $\boldsymbol{\beta_0}$ - החותך, $\boldsymbol{\beta_i}$ - השיפוע
- כמו ברל"פ, המקדמים β_i לא ידועים ומייצרים אומדי ריבועים פחותים b_i

בדיקת מובהקות מודל הרגרסיה השלם – מבחן *F*

מערכת ההשערות: $H_0: \beta_i = 0$, $H_1: else$ (מספיק ש-β_i אחד שונה מ-0 למובהקות)

סטטיסטי המבחן: $F = \frac{MSR}{MSE}$, **דחה אם:** $F_0 > F_{1-\alpha}^{(k,n-k-1)}$

הסקה על מקדמי הרגרסיה

1. <u>מבחן t:</u> זהה למבחן ברל"פ. מאפשר בחינה של כל פרמטר בנפרד. אם המשתנה ה-j שווה ל-0, אין למשתנה ה-j השפעה לינארית על המשתנה המוסבר.
$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$
סטטיסטי המבחן: $t_0 = \frac{b_j}{s(b_j)}$, **דחה אם** $\lvert t_0 \rvert > t_{1-\frac{\alpha}{2}}^{(n-k-1)}$
$s(b_j)$ - סטיית התקן של האומד למשתנה ה-j (מופיע בפלט הרגרסיה).
רב"ס דו"צ לפרמטר $\boldsymbol{\beta_j}$ ברמת סמך $\boldsymbol{1 - \alpha}$: $\beta_j \in \left[b_j \pm t_{1-\frac{\alpha}{2}}^{(n-k-1)} \cdot S(b_j) \right]$
2. <u>מבחן F חלקי:</u> בוחן את התרומה של קב' משתנים מסבירים להסבר הרגרסיה (פרמטר יחיד או קבוצת פרמטרים). משווים את SSR בלי קבוצת המשתנים לעומת הערך של רגרסיה שכוללת אותם.
סטטיסטי המבחן:

$$F_0 = \frac{(SSR_{full} - SSR_{partial})/(df_{full} - df_{partial})}{SSE_{full}/n - df_{full} - 1} = \frac{(R^2_{full} - R^2_{partial})/(df_{full} - df_{partial})}{(1 - R^2_{full})/n - df_{full} - 1}$$

דחה את H_0 בר"מ α אם: $F_0 > F_{1-\alpha}^{(df_{full} - df_{partial}, n - df_{full} - 1)}$

מקדם ההסבר ברגרסיה לינארית מרובה

מדד חלופי ל-R^2: $R^2_{adj.} \equiv 1 - \frac{SSE/n-k-1}{SST/n-1}$

n – מס' תצפיות, k – מס' משתנים מסבירים.

תמיד מתקיים $R^2_{adj} \leq R^2$. יש פחות תמריץ להוספת משתנים רבים למודל.

<u>**מולטיקולינאריות:**</u> כאשר קיים מתאם לינארי חזק בין משתנים מסבירים

1. **האומדים לשונות של מקדמי הרגרסיה (s_{b_i}) "מתנפחים":** הרב"סים למקדמים האמיתיים (β_i) הרבה יותר רחבים מכפי שאמור להיות.
2. **תוצאות מוזרות: סימני המקדמים הפוכים מהצפוי.**
3. **תוצאות מוזרות: מבחן F מעיד על מודל מובהק ומבחני t אינם מובהקים.**

מבחנים למולטיקולינאריות

1. **בדיקת מתאם בין כל זוג משתנים מסבירים:** $r_{x_j,x_k} = \frac{SS_{x_j,x_k}}{\sqrt{SS_{x_j} \cdot SS_{x_k}}}$
אם המתאם גבוה (קרוב ל-1 בערך מוחלט) כדאי לוותר על אחד מהמשתנים.
2. **מדד VIF:** לכל משתנה מסביר x_j מחשבים את הערך $VIF_j = \frac{1}{1 - R_j^2}$,
R_j^2 הוא מקדם ההסבר ברגרסיה בה המשתנה המוסבר הוא x_j והמשתנים המסבירים הם שאר המשתנים המסבירים. $\boldsymbol{VIF_j \geq 5}$ – יש מולטיקולינאריות

Regression Statistics	
Multiple R	$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{SS_x SS_y}} = \sqrt{R^2}$
R Square	$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{b_1^2 SS_x}{SS_y} = \frac{b_1^2 SS_x (\sum_{x_i} x_i^2 - n\bar{x}^2)}{\sum_{y_i} y_i^2 - n\bar{y}^2}$
Adjusted R Square	לא נדרש ברגרסיה פשוטה.
Standard Error	$S = \sqrt{S^2} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{SS_y - b_1^2 SS_x}{n-2}} = \sqrt{\frac{SSE}{n-2}}$
Observations	n

מה שמודגש נכון רק לרגרסיה לינארית פשוטה

אינטגרציה בחלקים - $\int u \cdot v' dx = u \cdot v - \int u' \cdot v dx$

<u>**אריתמטיקה של נגזרות:**</u>

$$(f \cdot g)'(x) = f'(x) \cdot g(x) + g'(x) \cdot f(x)$$

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - g'(x)f(x)}{g^2(x)}$$

זהו P-value בשימוש במבחן F.

ANOVA					
	(דרגות חופש) *Df-degrees of freedom*	*SS*	*MS*	*F*	*Significance F*
Regression	k – מס' המשתנים המסבירים (בפשוטה k=1)	SSR	$MSR = \frac{SSR}{k}$	$F_{stat} = \frac{MSR}{MSE}$	$Pv = P(F_{stat} < F_{cr})$ $F_{cr} = F_{1-\alpha}^{k,n-k-1}$
Residual	n-k-1	SSE	$MSE = \frac{SSE}{n-k-1} = S^2$		
Total	n-1	SST			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	$b_0 = \bar{y} - b_1 \bar{x}$	$S_{b_0} = S\sqrt{\left[\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right]}$	$T_{b_0} = \frac{b_0}{S_{b_0}}$	$P_{value} = P(\lvert T_{b_0} \rvert < T_{cr})$ $T_{cr} = t_{1-\frac{\alpha}{2}}^{n-2}$	$b_0 - t_{1-\frac{\alpha}{2}}^{n-2} S_{b_0}$	$b_0 + t_{1-\frac{\alpha}{2}}^{n-2} S_{b_0}$
X	$b_1 = \frac{SS_{xy}}{SS_x}$	$S_{b_1} = S/\sqrt{SS_x}$	$T_{b_1} = \frac{b_1}{S_{b_1}}$	$P_{value} = P(\lvert T_{b_1} \rvert < T_{cr})$ $T_{cr} = t_{1-\frac{\alpha}{2}}^{n-2}$	$b_1 - t_{1-\frac{\alpha}{2}}^{n-2} S_{b_1}$	$b_1 + t_{1-\frac{\alpha}{2}}^{n-2} S_{b_1}$

זהו ערך סטטיסטי המבחן במבחן t של ההשערה $H_0: \beta_1 = 0$ כנגד $H_0: \beta_1 \neq 0$.
ברגרסיה פשוטה בלבד, מתקיים $T_{b_1} = \sqrt{F_{stat}}$

זהו P-value של ההשערה $H_0: \beta_1 = 0$ כנגד $H_0: \beta_1 \neq 0$.
עבור רגרסיה פשוטה בלבד, זה שווה ל-P-value בשימוש במבחן F.

התפלגויות בדידות מיוחדות (כאשר p ההסתברות להצלחה בכל ניסיון)

התפלגות	אחידה	בינומית	היפר גיאומטרית	גיאומטרית	בינומית שלילית	פואסונית
תיאור	מ״מ בקפיצות של 1 עם הסתברות שווה לכל תוצאה	מספר ההצלחות בסדרת n ניסויי ברנולי ב״ת	מספר המיוחדים במדגם בגודל n (ללא החזרה) מאוכלוסיה בגודל N, שמתוכה D מיוחדים	מספר הניסיונות עד להצלחה הראשונה בסדרת ניסויי ברנולי ב״ת	מספר הניסיונות עד להצלחה ה- r בסדרת ניסויי ברנולי ב״ת	מספר האירועים ביחידת זמן נתונה
סימון	$X \sim \mathrm{U}(a,b)$	$X \sim \mathrm{B}(n,p)$	$X \sim \mathrm{HG}(N,D,n)$	$X \sim \mathrm{G}(p)$	$X \sim \mathrm{NB}(r,p)$	$X \sim \mathrm{P}(\lambda)$
$\mathrm{P}(X=k)$	$\begin{cases} \frac{1}{b-a+1}, & k=a,a+1,\dots,b \\ 0, & \text{אחרת} \end{cases}$	$\binom{n}{k} p^k q^{n-k}$ $k=0,1,\dots,n$	$\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}}$ $k=\max\{0,n-(N-D)\},\dots,\min\{n,D\}$	$q^{k-1}p$ $k=1,2,\dots$	$\binom{k-1}{r-1} p^r q^{k-r}$ $k=r,r+1,\dots$	$\frac{e^{-\lambda}\lambda^k}{k!}$ $k=0,1,\dots$
$\mathrm{E}(X)$	$\frac{b+a}{2}$	np	$\frac{nD}{N}$	$\frac{1}{p}$	$\frac{r}{p}$	λ
$\mathrm{V}(X)$	$\frac{(b-a+1)^2-1}{12}$	npq	$\frac{nD}{N}\left(1-\frac{D}{N}\right)\frac{N-n}{N-1}$	$\frac{q}{p^2}$	$\frac{rq}{p^2}$	λ
תכונות	עבור המקרה הפרטי $X \sim \mathrm{U}(a,b)$ $a=1, b=N$ נקבל: $P(X)=\frac{1}{N}$	<u>חיבור מ"מ בינומיים</u>: $X \sim B(n,p), Y \sim B(m,p)$ אותו p אז $X+Y \sim B(n+m,p)$		$\mathrm{P}(X>k)=q^k\ ;\ \mathrm{P}(X \ge k)=q^{k-1}$ $\mathrm{P}(X>a+b \mid X>a)=\mathrm{P}(X>b)$ $\mathrm{P}(X=a+b \mid X>a)=\mathrm{P}(X=b)$	<u>חשוב</u>: סכום של r מ"מ ב"ת שווי התפלגות המתפלגים גיאומטרית, מתפלג בינומי שלילי עם פרמטרים r,p	מספר האירועים בקטעי זמן זרים ב״ת $X \sim P(\lambda_1), Y \sim P(\lambda_2)$ $\Rightarrow X+Y \sim P(\lambda_1+\lambda_2)$

התפלגויות רציפות מיוחדות

התפלגות	אחידה	מעריכית	נורמלית
תיאור	מ״מ שיכול לקבל כל ערך בהסתברות שווה	א. אם מספר האירועים ביחידת זמן מתפלג פואסונית עם קצב אירועים λ, אז הזמן הבין מופעי (בין אירועים עוקבים) מתפלג מעריכית עם פרמטר λ. ב. אורך חיים של תהליכים המקיימים חוסר זיכרון	
סימון	$X \sim \mathrm{U}(a,b)$	$X \sim \exp(\lambda)$	$X \sim \mathrm{N}(\mu,\sigma^2)$
$f(x)$	$\begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{אחרת} \end{cases}$	$\begin{cases} \lambda e^{-\lambda x}, & 0<x \\ 0, & \text{אחרת} \end{cases}$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
$\mathrm{F}(x)$	$\begin{cases} 0, & x<a \\ \frac{x-a}{b-a}, & a \le x < b \\ 1, & b \le x \end{cases}$	$\begin{cases} 1-e^{-\lambda x}, & 0<x \\ 0, & \text{אחרת} \end{cases}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$
$\mathrm{E}(X)$	$\frac{b+a}{2}$	$\frac{1}{\lambda}$	μ
$\mathrm{V}(X)$	$\frac{(b-a)^2}{12}$	$\frac{1}{\lambda^2}$	σ^2
תכונות	במקרה הפרטי $X \sim U(0,1)$ נקבל: $F(x)=\begin{cases} 0, & x<a \\ x, & a \le x < b \\ 1, & b \le x \end{cases}$	$\mathrm{P}(X>a+b \mid X>a)=\mathrm{P}(X>b)$ אם מ״מ מקיים את תכונת חוסר הזיכרון הוא מתפלג אקספוננציאלית ולהיפך. דוגמא: אם λ ממוצע תאונות דרכים בחודש אז $X \sim P(\lambda)$ מספר התאונות בחודש, ו $X \sim \exp(\lambda)$ - זמן בין 2 תאונות. $P(X \le t)=1-e^{-\lambda t}$, $P(X>t)=e^{-\lambda t}$	$f(\mu+x)=f(\mu-x)$, $Z_{(1-\alpha)}=-Z_\alpha$ $Z=\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1^2)$ $\Phi(-b)=1-\Phi(b)$, $p(x \ge a)=1-p(x \le a)$ $\mathrm{P}(a<Z<b)=\Phi(b)-\Phi(a)$ $\mathrm{P}(-c<Z<c)=2\Phi(c)-1$