

# מושגים סטטיסטיים, חזרה (אולי בזווית קצת שונה)

# סטטיסטיקה

- עיבוד נתונים וארגונים
- הצגת נתונים, מדדים תיאוריים
- הסקה סטטיסטית:
  - אמידה נקודתית
  - רוחי סמך
  - מבחני השערה
- מודלים סטטיסטים

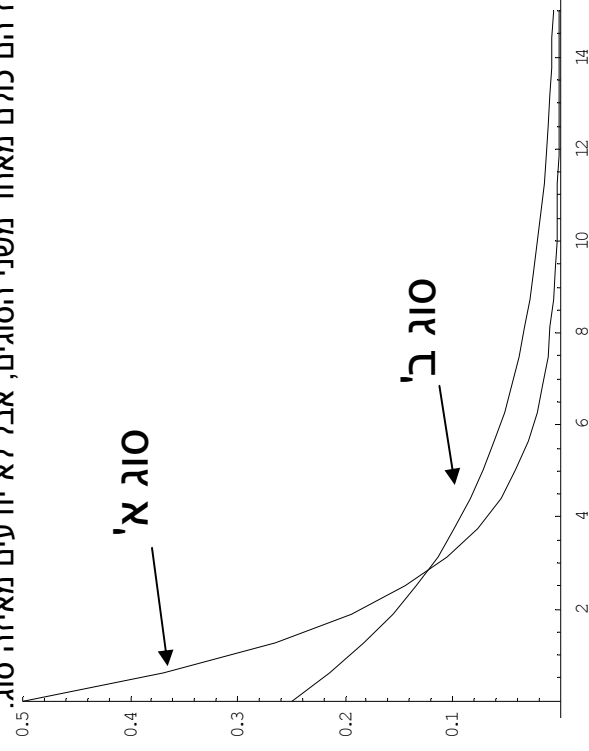
# הסתברות – כלי יסודי בהסקה סטטיסטית

- משתנים מקריים
- סטטיסטי הוא משתנה מקרי
- ידע לגבי פילוג הסטטיסטי מאפשר
  - למצוא (אמדים נקודתיים) סטטיסטיים:
    - חסרי הטיה.
    - עלי שונות קטנה.
  - ליצור רווחי סמך.
  - לתכנן מבחני השערה.

# דוגמא למבחן השערה: אורך חיי רכיב

$$f(x) = \lambda e^{-\lambda x}$$

- רכיבים רבים הינם בעלי אורך חיים (זמן מתחילת שימוש עד קלקול) בעלי התפלגות אקספוננציאלית.
- לידנו נופל ארגז רכיבים ואנו יודעים שהרכיבים בארגז הם כולם מאחד משני הסוגים, אבל לא יודעים מאיזה סוג:
  - סוג א, אורך חיים  $\exp(1/2)$
  - סוג ב, אורך חיים  $\exp(1/4)$



- אנו לוקחים רכיב בודד ( $n=1$ ), בודקים את אורך החיים שלו ומבצעים את מבחן ההשארה הבא:
  - HO: הרכיבים מסוג א'
  - H1: הרכיבים מסוג ב'
- סטטיסטי המבחן הוא אורך חיי הרכיב: X.
- חוק הדחייה (של H0), נדחה אם  $X > a$ .
- טעות מסוג ראשון, סוג שני ועוצמה:

$$\alpha = P(\text{reject} | H_0)$$

$$\beta = P(\text{accept} | H_1) \quad \text{Power} = 1 - \beta = P(\text{reject} | H_1)$$

מודלים סטטיסטים ב' ארתור צ'ירגייב, יוני נצרת

## רוצים $\alpha=0.05$ אז מהו $a$ ?

$$0.05 = \int_a^{\infty} \frac{1}{2} e^{-1/2x} dx = e^{-a/2}$$

$\Downarrow$

$$\ln(0.05) = -a/2$$

$\Downarrow$

$$a = 5.99$$

חישבנו כאן את האחוזון ה  $1-0.05=0.95$  של  $\exp(1/2)$ .

עבור ההתפלגויות שימושיות רבות (נורמאלית,  $t$ ,  $F$ ,  $\chi^2$ ) לא ניתן לפתור את האינטגרל לעיל ולכן לא ניתן לחשב באופן אנליטי את האחוזון  $\leq$  ולכן הטבלאות.

יש טבלאות עבור אחוזונים ויש טבלאות עבור שטחים.

## מהי עוצמת המבחן?

- רמת המובהקות הבטיחה לנו שהסיכוי שנדחה את  $H_0$  בשוגג הוא 0.05.
- עוצמת המבחן היא הסיכוי לדחות את  $H_0$  בצדק:

$$Power = 1 - \beta = P(reject | H_1) = \int_{5.99}^{\infty} \frac{1}{4} e^{-x/4} dx = e^{-5.99/4} = 0.22$$

שזה לא כל כך טוב...

# תפיסת היישום של מבחני השערה

■  $H_0$  לרוב מייצג את המצב הקיים.

■  $H_1$  מצב חדש:

□ בביוסטטיסטיקה, שיפור הנגרם ע"י תרופה.

□ במדעים, תיאוריה מדעית חדשה.

■ לדוגמא ברגרסיה ליניארית פשוטה כאשר מבצעים

מבחן  $F$ , אז  $H_0$  טוענת שהשיפוע הוא אפס (אין

השפעה של המשתנה  $X$  על  $Y$ ) ו $H_1$  אחרת (יש

השפעה).

## P-Value

- נניח וביצענו את המבחן ויצא  $X=12.3$ .
- עבור מבוהקות  $0.05$  צריך לדחות כי  $5.99 < 12.3$
- מהי רמת המובהקות המינימלית שבה נדחה:

$$PV = \int_{12.3}^{\infty} \frac{1}{2} e^{-x/2} dx = e^{-12.3/2} = 0.002$$

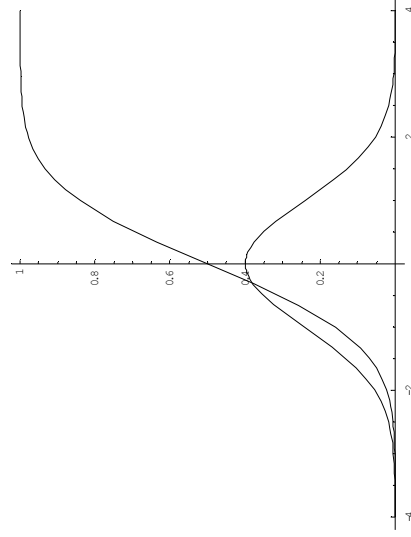


# ההתפלגות הנורמאלית והחברים של...ה

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# ההתפלגות הנורמאלית...

- משפט הגבול המרכזי: סכום של אוסף משתנים מקריים i.i.d. מתפלג בקרוב נורמאלי.
  - מכאן ההתפלגות הנורמאלית היא כ"כ פופולריות.
  - ממוצע (חשבוני) הוא סכום כפול קבוע  $(1/n)$ .
- להתפלגות הנורמאלית תכונות מתמטיות "נוחות" הקשורות למודלים סטטיסטיים ליניאריים



$$X \sim N(\mu, \sigma^2)$$

פונ' הצפיפות ופונ' התפלגות מצטברת  
נורמאלית סטנדרטית:

# פילוג הממוצע....

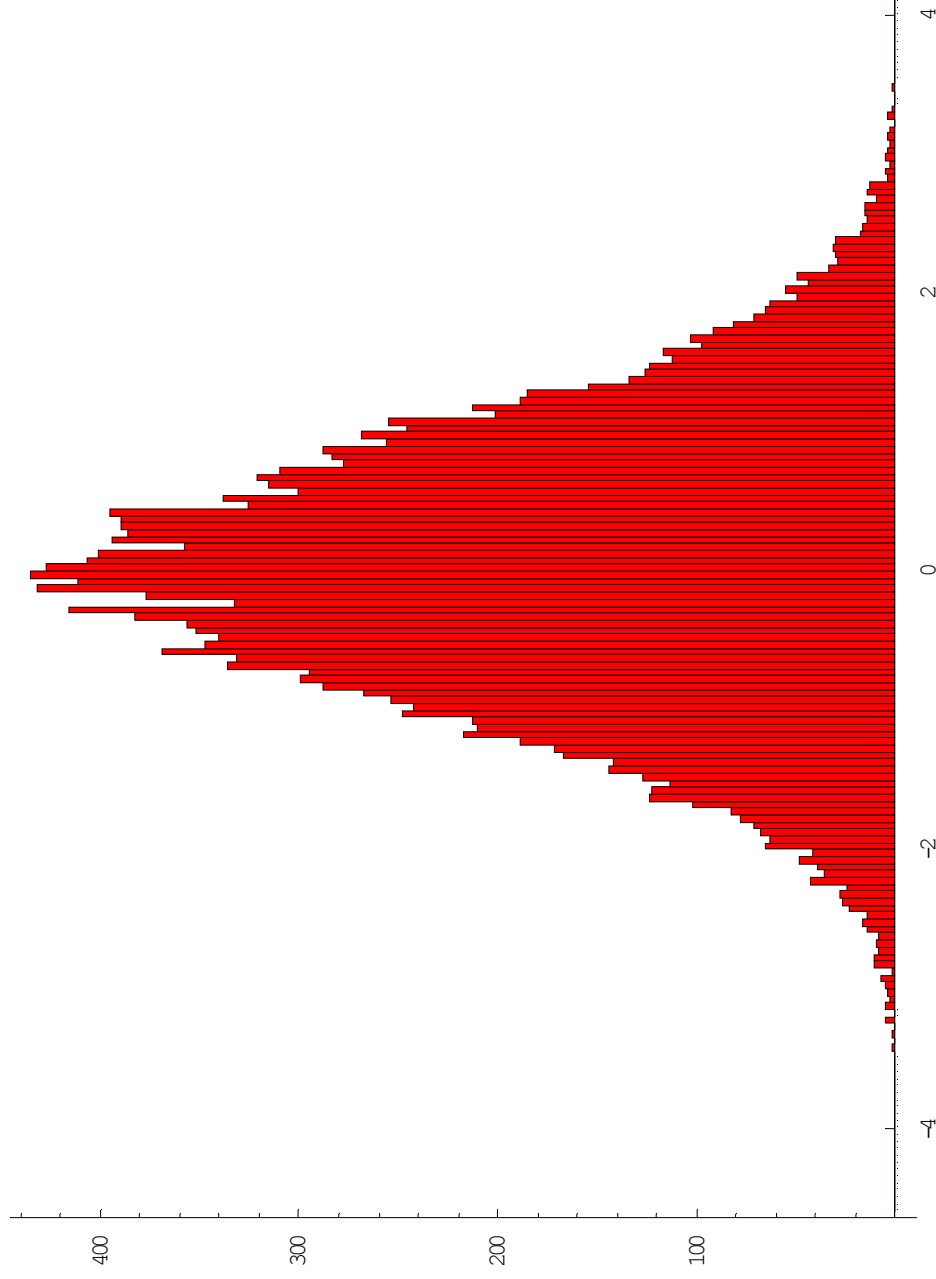
$$\begin{aligned}\bar{X} &= \frac{1}{n} \sum_{i=1}^n X_i \Rightarrow E\bar{X} = \frac{1}{n} n EX = EX \\ &\Downarrow \\ \text{Var}(\bar{X}) &= \frac{1}{n^2} n \text{Var}(X) = \frac{\text{Var}(X)}{n} \\ &\Downarrow \\ \text{SD}(\bar{X}) &= \frac{\text{SD}(X)}{\sqrt{n}}\end{aligned}$$

$$\begin{aligned}X_1 \sim N(\mu_1, \sigma_1^2), \quad \dots, \quad X_n \sim N(\mu_n, \sigma_n^2), \quad \text{independent} \\ &\Downarrow \\ \sum_{i=1}^n a_i X_i &\sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right) \\ &\Downarrow \\ X_i - X_j &\sim N(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)\end{aligned}$$

$$\begin{aligned}X_1, \dots, X_n &\sim \text{NID}(\mu, \sigma^2) \\ &\Downarrow \\ \bar{X} &\sim N\left(\mu, \frac{\sigma^2}{n}\right)\end{aligned}$$

# סימולציה מחשב של משתנים מקריים $N(0,1)$

```
ivals = RandomArray[ndist, {20000, 7}];
```



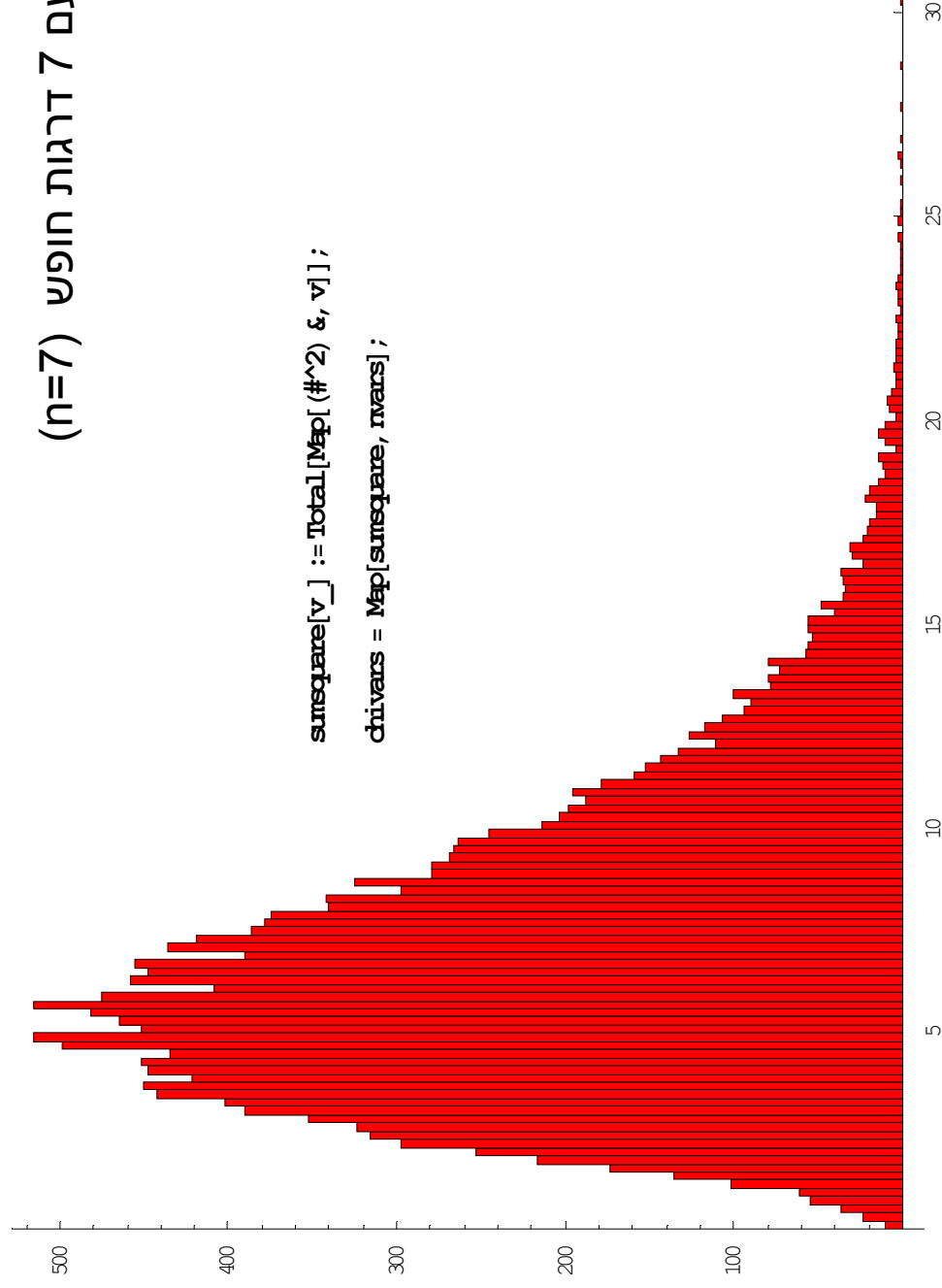
ח: סכום של ריבועי נורמאליים סטנדרטיים

$$Z_i \sim NID(0, 1)$$

$$\sum_{i=1}^n Z_i^2 \sim \chi_n^2$$

# חי: סכום של ריבועי מ"מ נורמאליים סטנדרטים

התפלגות חי עם 7 דרגות חופש (n=7)



מודלים סטטיסטיים ב' ארתור צ'ירגיני, יוני נצרת

# פילוג השונות המדגמית...

עבור מדגם מקרי מהתפלגות נורמאלית:

$$(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

ומכאן התפלגות חי שימושית עבור הסקה לגבי השונות.  
לדוגמא, רוח סמך עבור השונות:

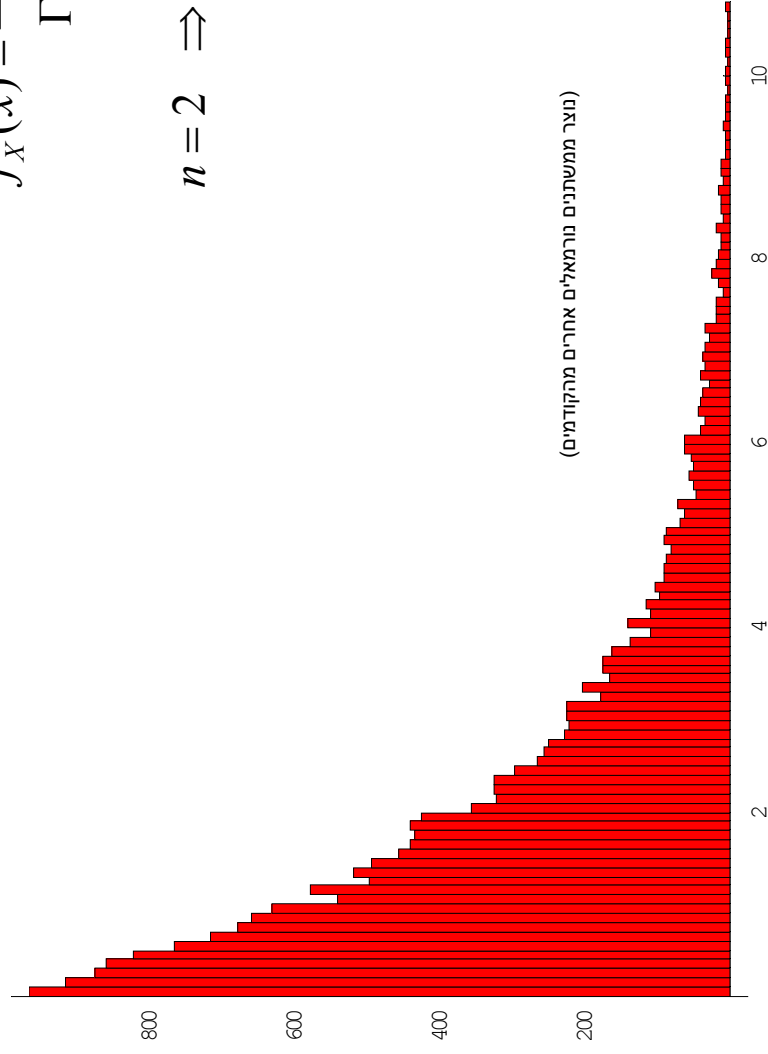
$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2, n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2, n-1}}\right) = 1 - \alpha$$

מודלים סטטיסטים ב' ארתור צ'ירגייב, יוני נצרת

# חי עם 2 דרגות חופש = $\exp(1/2)$

$$f_X(x) = \frac{2^{-\frac{n}{2}}}{\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

$$n=2 \Rightarrow f_X(x) = \frac{2^{-1}}{\Gamma(1)} e^{-\frac{x}{2}} = \frac{1}{2} e^{-\frac{x}{2}}$$



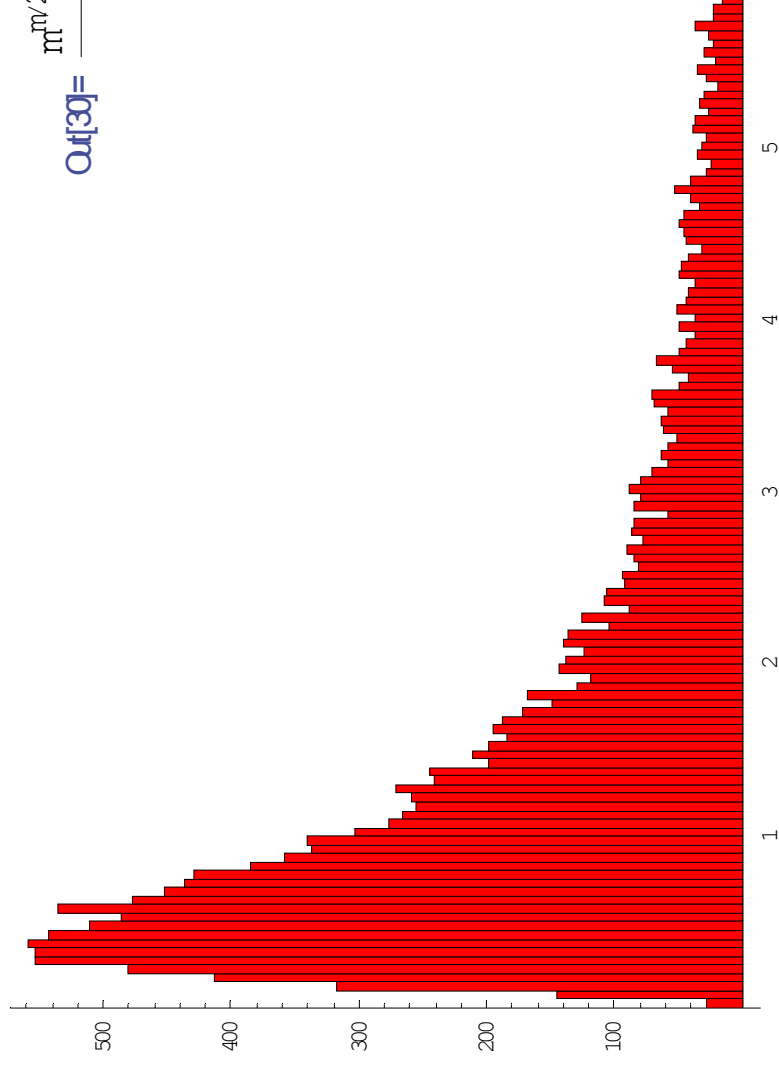


# F: מנה של 2 חי בדרגות החופש.

$$fvars = \frac{chivars/7}{chivars2/2};$$

In[30]: **EDF**[ERatioDistribution[n, m], x]

$$Out[30]= \frac{m^{m/2} n^{n/2} x^{-1+\frac{n}{2}} (m+nx)^{\frac{1}{2}(-m-n)}}{\text{Beta}\left[\frac{n}{2}, \frac{m}{2}\right]}$$



מודלים סטטיסטים ב' ארתור צ'ירגייב, יוני נצרת

## מדוע F שימושית.

- ראינו שסכומי ריבועים מתפלגים ח.<sup>1</sup>
- לכן מנות של סכומי ריבועים מתפלגים F.
- במבוא לסטטיסטיקה, השימוש היה לצורך הסקה לגבי המנה של שוניות של אוכלוסיות.
- בניתוח שונות השימוש יהיה השווה של MSE... נראה בהמשך.

# הצורך בהתפלגות t...

- כאשר מבצעים הסקה לגבי תוחלת אוכלוסיה עם שונות ידועה משתמשים בפילוג של  $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$  (נורמאלי סטנדרטי).

לבנות רווחי סמך ומבחני השארה.

- לרוב השונות כמובן אינה ידועה... ולכן הגיוני להחליף את  $\sigma$  ב  $S$ . מכאן אנו מתעניינים בפילוג של  $\frac{\bar{X} - \mu}{S / \sqrt{n}}$

- כאשר המדגם גדול (לדוגמא  $n > 120$ ) אז  $S$  אומד את  $\sigma$  באופן מאוד מדויק ולכן  $\frac{\bar{X} - \mu}{S / \sqrt{n}}$  מתפלג בקירוב נורמאלי סטנדרטי.

- עבור מדגמים יותר קטנים אנו מתעניינים בפילוג המדויק של  $\frac{\bar{X} - \mu}{S / \sqrt{n}}$  ..... התפלגות t עם n-1 דרגות חופש....

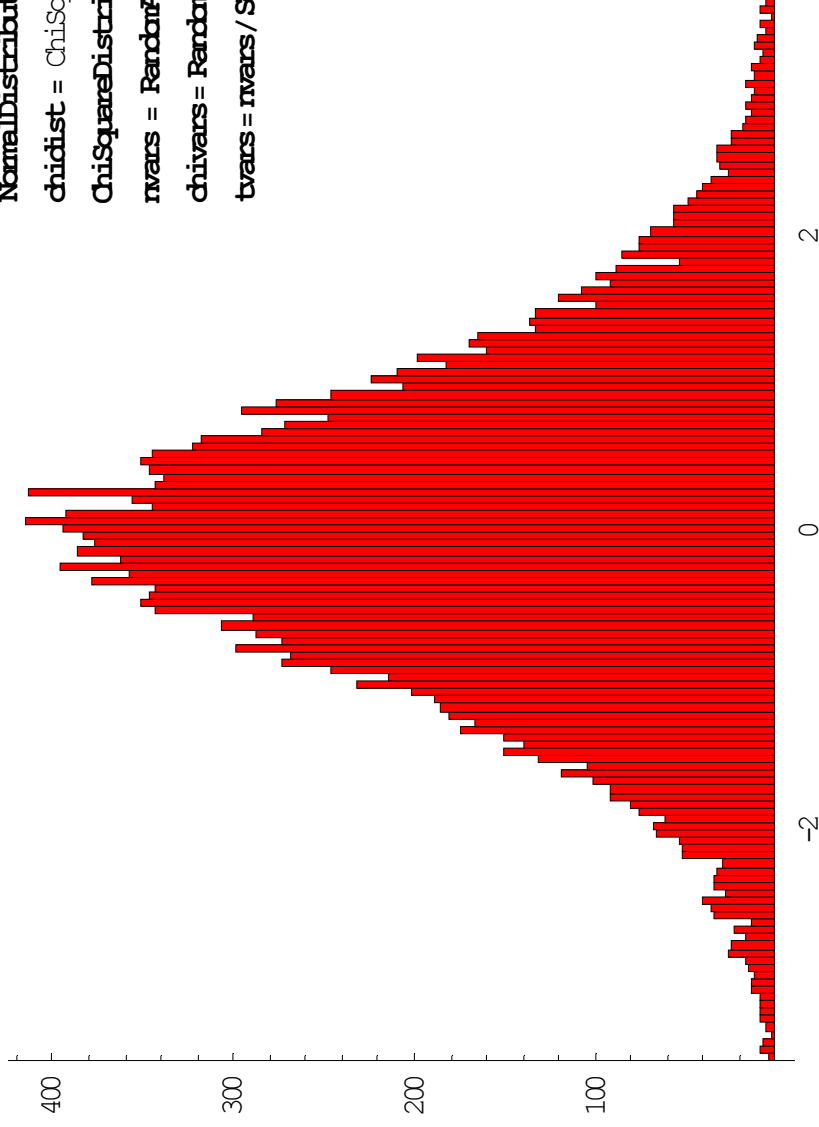
- נכפיל ונחלק ב  $\sigma$  ונקבל:  
$$\frac{(\bar{X} - \mu) / (\sigma / \sqrt{n})}{\sqrt{S^2 / \sigma^2}}$$

- אם כך המונה מתפלג נורמאלי סטנדרטי והמכנה מתפלג כמו  $\sqrt{\chi_{n-1}^2 / (n-1)}$
- תוצאה חשובה וממעניינת נוספת היא שהשונות המדגמית ב"ת בממוצע המדגמי.
- ומכאן התפלגות t: מנה של מ"מ נורמאלי סטנדרטי ב מ"מ  $\sqrt{\chi_{n-1}^2 / (n-1)}$  ב"ת.

מודלים סטטיסטים ב' ארתור צ'ירגייב, יוני נצרת

# סימולציה של מ"מ t

```
rndist = NormalDistribution[0, 1]
NormalDistribution[0, 1]
chidist = ChiSquareDistribution[7]
ChiSquareDistribution[7]
rwars = RandomArray[rndist, {20000}];
chivers = RandomArray[chidist, {20000}];
twars = rwars / Sqrt[chivers / 7];
```



In[3]= **EDF**[StudentTDistribution[n], x]

$$\text{Out[3]} = \frac{\left(\frac{n}{n+x^2}\right)^{\frac{1+n}{2}}}{\sqrt{n} \text{Beta}\left[\frac{n}{2}, \frac{1}{2}\right]}$$

# הקשר בין $t$ ל $F$

$$t = \frac{Z}{\sqrt{\chi_{n-1}^2 / (n-1)}}$$

$$t^2 = \frac{Z^2}{\chi_{n-1}^2 / (n-1)} = \frac{\chi_1^2 / 1}{\chi_{n-1}^2 / (n-1)} = F_{1, n-1}$$

מכאן גם הקשר בין מבחן  $t$  להשוואת 2 אוכלוסיות לבין ניתוח שונות...  
נראה בשבועות הקרובים.