מודלים סטטיסטיים

ד"ר מאיר קרלינסקי

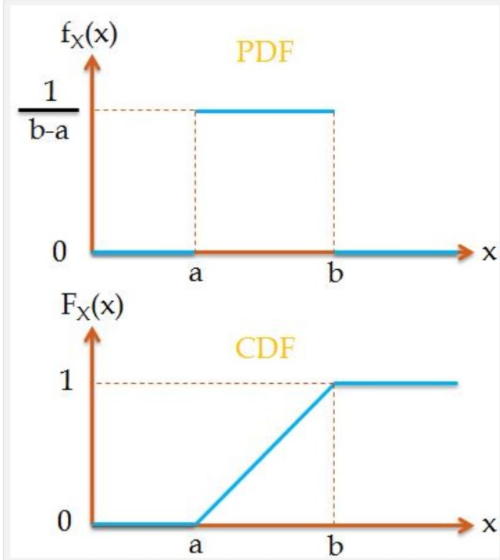
6. תורת ההסתברות (3)

- משתנים מקריים רציפים:
  - אחיד - Uniform
  - מעריכי - Exponential
  - נורמלי - Normal
  - ונגזרותיו: חי-בריבוע - Chi-square
    - טי (סטודנט) - t (Student)
    - אף (סנדקור) - F (Snedecor)
- חוקי גבול: אי-שיוויוני מרקוב וצ'בישב
  - חוקי המספרים הגדולים
  - חוק הגבול המרכזי

התפלגויות משתנים מקריים - רציפים:

מ"מ אחיד (רציף) – Uniform X~ U(a , b)

$$X = \begin{cases} a \le x \le b & f(x) = 1/(b-a) \\ \text{אחרת} & f(x) = 0 \end{cases}$$

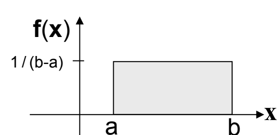


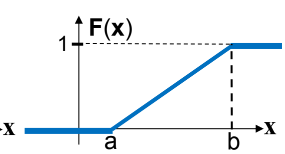
PDF and CDF of Continuous Uniform Distribution (Continous)

התפלגויות משתנים מקריים - רציפים:

מ"מ אחיד (רציף) – Uniform X~ U(a , b)

$$X = \begin{cases} a \le x \le b & f(x) = 1/(b-a) \\ \text{אחרת} & f(x) = 0 \end{cases}$$





תוחלת מ"מ אחיד (רציף)

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\,dx = \int_a^b x \times \frac{1}{b-a}\,dx =$$

$$= \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left[\frac{1}{2}x^2\right]_a^b = \frac{1}{b-a}\left[\frac{1}{2}b^2 - \frac{1}{2}a^2\right] = \frac{b+a}{2} =$$

נק' האמצע
בין a ל- b

שונות מ"מ אחיד (רציף)

$$\sigma_X^2 = Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2\, f_X(x)\,dx =$$

$$= \int_a^b \left[x - \frac{b+a}{2}\right]^2 \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b \left[x^2 - 2\times\frac{b+a}{2}\times x + \left(\frac{b+a}{2}\right)^2\right]dx =$$

$$= \frac{1}{b-a}\left[\int_a^b x^2\,dx - (b+a)\int_a^b x\,dx + \int_a^b \left(\frac{b+a}{2}\right)^2 dx\right] = \frac{(b-a)^2}{12}$$

התפלגויות משתנים מקריים - רציפים:

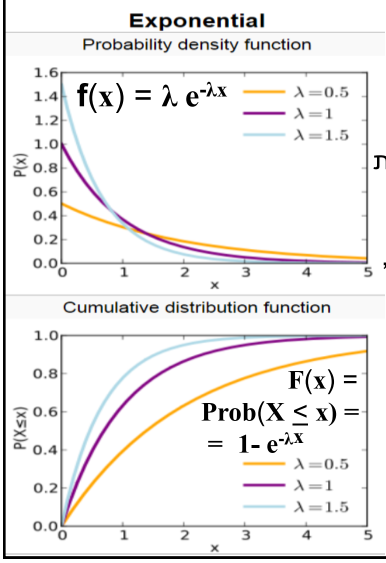
ללב

$X\sim Exp(\lambda)$ מ"מ מעריכי – Exponential

$$X = \begin{cases} 0 \le x & f(x) = \lambda e^{-\lambda x} \quad \lambda > 0 \\ x < 0 & f(x) = 0 \end{cases}$$

מתאר את אורך הרווח / הזמן בין מופעי מ"מ פואסוני.
מהווה (במובן מסוים) אנאלוגיה רציפה למ"מ גיאומטרי
וכמוהו הינו (היחיד בין המ"מ-ים הרציפים) **חסר-זיכרון:**

$$\mathrm{Prob}(X > s+t \mid X > s) = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \mathrm{Prob}(X > t)$$



כמ"מ הקשור למ"מ פואסוני, משמש בתיאור אותן תופעות
ומתאים לתיאור הזמן בין אירועים המתרחשים באקראי
אך בקצב ממוצע קבוע: זמני חיים של חומר רדיו-אקטיבי
(ובקביעת 'זמן מחצית החיים' = חציון המ"מ: $\ln 2/\lambda$),
זמן עד תקלה במכשיר, זמן עד הופעת שיחה בקו טלפון,
זמן עד הופעת לקוח, זמני נתינת שירות, וכד'.

תוחלת מ"מ מעריכי

$$E(X) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_0^{\infty} x\, \lambda e^{-\lambda x}\, dx = 1/\lambda$$

תוחלת מ"מ מעריכי

$$\sigma_X^2 = Var(X) = \int_{-\infty}^{\infty} [x - E(X)]^2 f_X(x)\, dx = \int_0^{\infty} [x - 1/\lambda]^2 \lambda e^{-\lambda x} dx = 1/\lambda^2$$

הקשר בין מ"מ מעריכי למ"מ פואסוני $X\sim Exp(\lambda)$ $Y\sim Poisson(\lambda)$ **ללב**

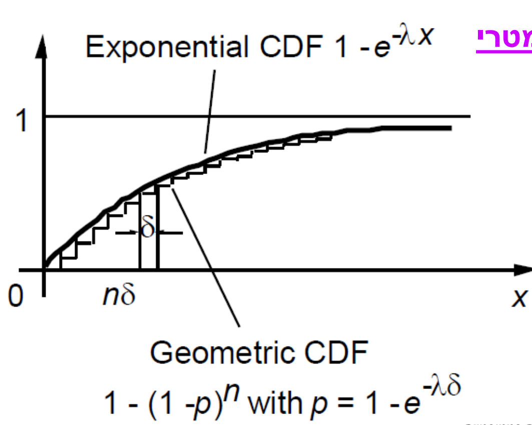
בתהליך (בזרם אירועים) פואסוני בעל קצב λ, משך הזמן עד התרחשות האירוע הראשון
(או משך הזמן בין שני אירועים סמוכים/עוקבים) הינו מ"מ מעריכי עם פרמטר λ "

לדוגמא: לומר "למרכזיה מגיעות שיחות בקצב פואסוני של 20 שיחות בשעה" שקול למשפט:
"הפרמטר של ההתפלגות המעריכית הרלוונטית הינו 20/60 = 1/3 = λ שיחות בדקה.
לחילופין ניתן להגיד " התפלגות מעריכית עם ממוצע/תוחלת של 3 [= (1/3) / 1 = $1/\lambda$].

הקשר בין מ"מ-ים מעריכי וגיאומטרי

כאמור, מ"מ מעריכי (אקספוננציאלי)
מהווה אנאלוגיה רציפה למ"מ גיאומטרי.
ניתן לראות זאת בגרף משמאל
שבו פונקציות ההתפלגויות המצטברות
של מ"מ-ים אילו מקיימות:

$$F^{\exp}(n\delta) = F^{\mathrm{geo}}(n), \qquad n = 1, 2, \ldots,$$



כאשר האינטרוול δ הינו כזה ש:

$$e^{-\lambda\delta} = 1 - p.$$

וכש- δ שואף ל- 0 המ"מ המעריכי
הינו הגבול של המ"מ הגיאומטרי.

טרנספורמציה ליניארית של משתנה מקרי

If X is continuous, the PDF and the CDF can be obtained from each other by integration or differentiation:

פונקציית ההתפלגות המצטברת - CDF
$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt,$$

פונקציית צפיפות ההסתברות - PDF
$$f_X(x) = \frac{dF_X}{dx}(x).$$

The PDF of a Linear Function of a Random Variable

Let X be a continuous random variable with PDF f_X, and let

$$Y = aX + b,$$

for some scalars $a \neq 0$ and b. Then,

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

טרנספורמציה ליניארית של משתנה מקרי (המשך 1)

The PDF of a Linear Function of a Random Variable

Let X be a continuous random variable with PDF f_X, and let

$$Y = aX + b,$$

for some scalars $a \neq 0$ and b. Then,

פונקציית צפיפות ההסתברות - PDF
$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

הוכחה (ללב)

steps for the case where $a > 0$; the case $a < 0$ is similar. We have

$$\begin{aligned} F_Y(y) &= \mathbf{P}(Y \leq y) \\ &= \mathbf{P}(aX + b \leq y) \\ &= \mathbf{P}\left(X \leq \frac{y-b}{a}\right) \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

We now differentiate this equality and use the chain rule, to obtain

$$f_Y(y) = \frac{dF_Y}{dy}(y) = \frac{1}{a} \cdot \frac{dF_X}{dx}\left(\frac{y-b}{a}\right) = \frac{1}{a} \cdot f_X\left(\frac{y-b}{a}\right)$$

טרנספורמציה ליניארית של מ"מ (המשך 2)

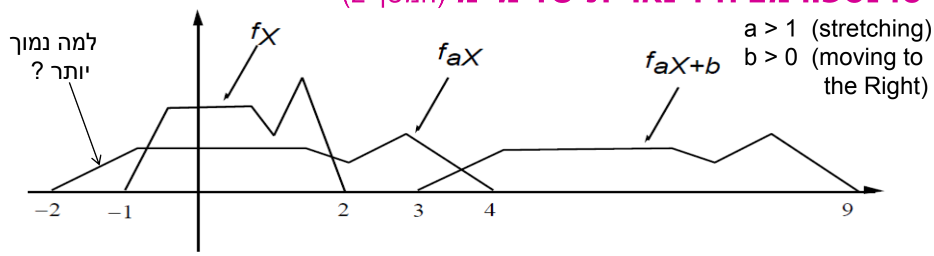


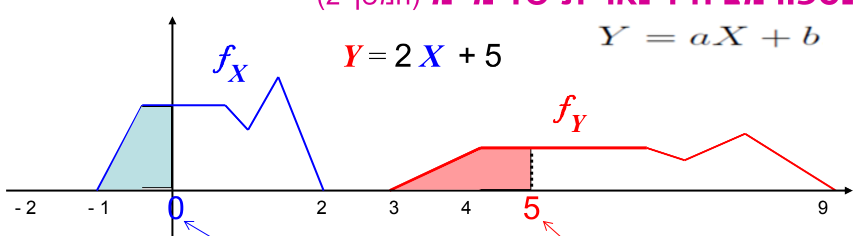
Figure 3.21: The PDF of $aX + b$ in terms of the PDF of X. In this figure, $a = 2$ and $b = 5$. As a first step, we obtain the PDF of aX. The range of Y is wider than the range of X, by a factor of a. Thus, the PDF f_X must be stretched (scaled horizontally) by this factor. But in order to keep the total area under the PDF equal to 1, we need to scale the PDF (vertically) by the same factor a. The random variable $aX + b$ is the same as aX except that its values are shifted by b. Accordingly, we take the PDF of aX and shift it (horizontally) by b. The end result of these operations is the PDF of $Y = aX + b$ and is given mathematically by

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

If a were negative, the procedure would be the same except that the PDF of X would first need to be reflected around the vertical axis ("flipped") yielding f_{-X}. Then a horizontal and vertical scaling (by a factor of $|a|$ and $1/|a|$, respectively) yields the PDF of $-|a|X = aX$. Finally, a horizontal shift of b would again yield the PDF of $aX + b$.

טרנספורמציה ליניארית של מ"מ (המשך 2)

$Y = 2X + 5$ $\qquad$ $Y = aX + b$



$X = 0 \;\rightarrow\; Y = 2 \times 0 + 5 = 5$

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \qquad F_Y(y) = F_X\left(\frac{y-b}{a}\right)$$

$f_Y(Y=5) = ½ f_X[\,(5 - 5)/2\,] =$
$= ½ f_X(0)$

$F_Y(Y=5) = F_X[\,(5 - 5)/2\,] =$
$= F_X(0)$

Standardized Random Variable - משתנה מקרי מתוקנן

If X is a random variable with mean, μ, and standard deviation, σ, we define the corresponding **standardised random variable** as:

$$Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$$

$$E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)f(x)dx = \frac{1}{\sigma}\left[\int_{-\infty}^{\infty} xf(x)dx - \mu\int_{-\infty}^{\infty} f(x)dx\right] = \frac{1}{\sigma}(\mu-\mu) = 0$$

$$Var\left(\frac{X-\mu}{\sigma}\right) = \int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}-0\right)^2 f(x)dx = \frac{1}{\sigma^2}\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \frac{1}{\sigma^2}Var(X) = \frac{\sigma^2}{\sigma^2} = 1$$

למשתנה מקרי **מתוקנן** **תוחלת** = **0** ו**שונות** (וגם **סטיית-תקן**) = **1**

התפלגויות משתנים מקריים - רציפים:

מ"מ נורמלי
Normal Distribution

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

ההתפלגות הנורמלית היא השימושית ביותר בסטטיסטיקה - תיאורטית וביישומיה בכל התחומים.
תועלתה הרבה בסטטיסטיקה היסקית (אמידה ובדיקת השערות) נובעת הן **ממשפט הגבול המרכזי**, לפיו הסכום או הממוצע של משתנים מקריים ב"ת ובעלי אותה התפלגות, לאחר תיקנון מתאים, מתכנס בהתפלגות אל ההתפלגות הנורמלית, והן מהיותה המקור להתפלגויות שימושיות נוספות: F , t , χ^2

ההתפלגות הנורמלית הסטנדרטית (קרויה גם התפלגות **Z**), בעלת תוחלת 0, ושונות 1, מתועדת בפירוט רב בטבלאות (לחישוב הסתברויות וערכים קריטיים בהסקה) וניתן להגיע אליה בנקל מכל משתנה מקרי בעל התפלגות נורמלית כללית על ידי הזזה (הוספת קבוע) ומתיחה (הכפלה בקבוע)
זהות התפלגות נורמלית מסוימת נקבעת באופן חד-ערכי על-פי שני פרמטרים: התוחלת והשונות שלה.

ההתפלגות הנורמלית נקראת גם 'גאוסיאן' על שמו של קרל פרידריך גאוס, או 'עקומת הפעמון' על פי צורת גרף **פונקציית צפיפות ההסתברות** שלה:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} \quad -\infty < x < +\infty,$$

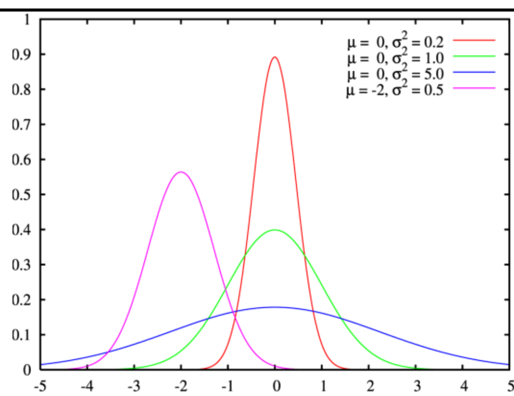
$$e = 2.7183 \; \ldots$$
$$\pi = 3.1416 \; \ldots$$

פונקציית ההתפלגות - CDF (ההסתברות) המצטברת :
(אין לה ביטוי אנליטי סגור מפורש)

$$\Phi(x) = \text{CDF}\;(x) = \int_{-\infty}^{x} f(t)\,\text{d}t = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{d}t$$

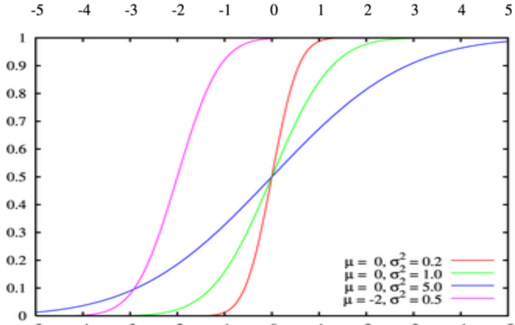
גרפי התפלגויות נורמליות

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



פונקציות צפיפות ההסתברות של התפלגויות נורמליות

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

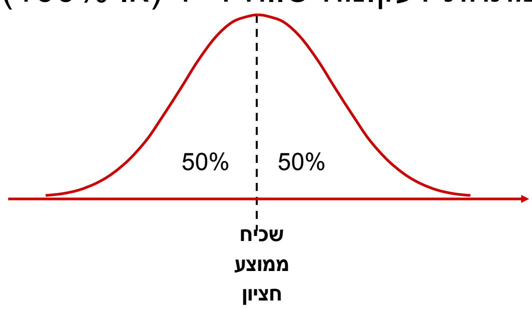


פונקציית ההתפלגות (ההסתברות) המצטברת - CDF של התפלגויות נורמליות

$$\Phi(x) = \text{CDF}\;(x) =$$
$$= \int_{-\infty}^{x} f(t)\,\text{d}t = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \text{d}t$$

איפיוני ההתפלגות הנורמלית:

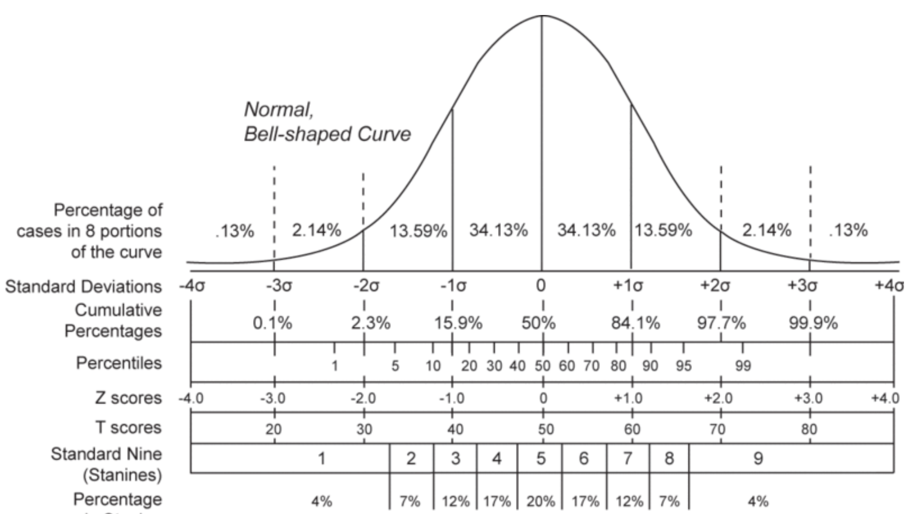
- מתאימה למשתנה רציף מסולם רווח (אינטרוולי) ומעלה.
- חד-שיאית (חד-מודלית)
- סימטרית
- מוגדרת על כל הישר הממשי ($-\infty < x < +\infty$)
- השכיח = הממוצע = החציון
- השטח שמתחת לעקומה שווה ל- 1 (או 100%)

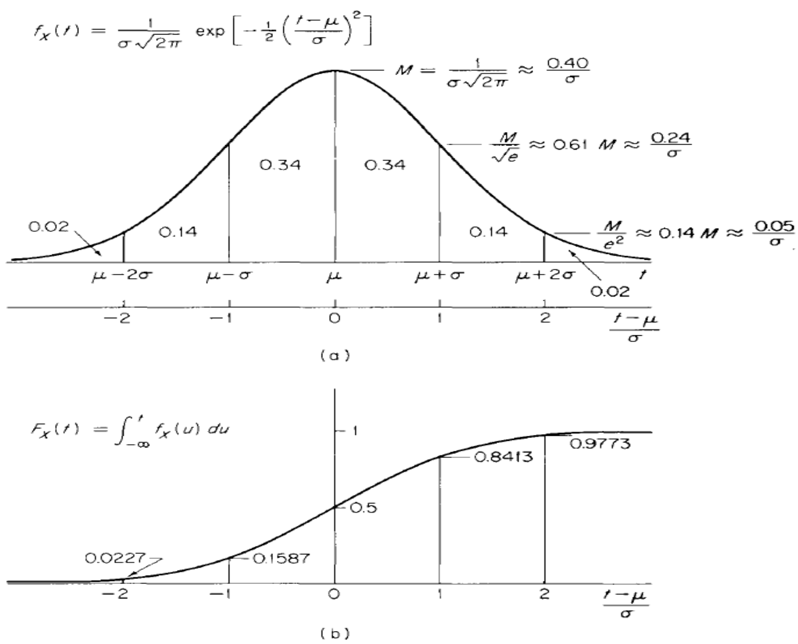


התפלגות נורמלית סטנדרטית

$$f(Z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{Z^2}{2}}$$

$$\mathbf{Z} \sim N(0\ ,\ 1)$$





Probability density and distribution functions for a normally distributed random variable.

(המשך) מ"מ נורמלי

תוחלת מ"מ נורמלי

$$E(X) = \int_{-\infty}^{+\infty} x f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x \;=\; \mu$$

שונות מ"מ נורמלי

$$Var(X) = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x \;=\; \sigma^2$$

כלומר, הפרמטרים μ , σ^2 המופיעים בביטוי לפונקציית צפיפות ההסתברות הם למעשה התוחלת והשונות (בהתאמה) של המ"מ שזו התפלגותו.

מכיון שאילו הם הפרמטרים היחידים, קביעתם (או ידיעתם) מגדירה באופן חד ערכי את התפלגות המ"מ הנורמלי - ומסמנים זאת כך: $\mathbf{X} \sim N(\mu\,,\,\sigma^2)$

<u>**למ"מ בעל התפלגות הנורמלית** יש כמה תכונות מועילות ושימושיות:</u>

טרנספורמציה לינארית שלו יוצרת מ"מ שהינו גם בעל התפלגות נורמלית עם הפרמטרים הבאים –

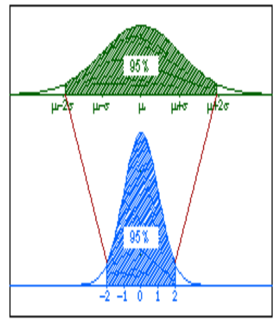
$$\mathbf{X} \sim N(\mu\,,\,\sigma^2) \quad \rightarrow \quad a\mathbf{X}+\mathbf{b} \;\sim\; N(a\mu + b\,,\,a^2\sigma^2)$$

ולכן **תיקנון** ע"י טרנספורמציה לינארית בה $a = 1/\sigma$, $b = -\mu/\sigma$ או $\mathbf{Z} = (\mathbf{X} - \mu)/\sigma$
מביא למ"מ נורמלי סטנדרטי - בעל התפלגות נורמלית עם תוחלת 0 ושונות 1: $\mathbf{Z} \sim N(\mu{=}0\,,\,\sigma^2{=}1)$
שכאמור, מתועדת בפירוט בטבלאות.

מ"מ נורמלי (המשך)

A linear function of a normal random variable is normal. Suppose that X is a normal random variable with mean μ and variance σ^2, and let $Y = aX + b$, where a and b are some scalars. We have

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



Therefore,

$$\begin{aligned} f_Y(y) &= \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \\ &= \frac{1}{|a|}\frac{1}{\sqrt{2\pi}\,\sigma} e^{-((y-b)/a)-\mu)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi}\,|a|\sigma} e^{-(y-b-a\mu)^2/2a^2\sigma^2}. \end{aligned}$$

We recognize this as a normal PDF with mean $a\mu + b$ and variance $a^2\sigma^2$. In particular, Y is a normal random variable.

התפלגות נורמלית סטנדרטית

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}$$

$\mathbf{Z} \sim N(0\ ,\ 1)$

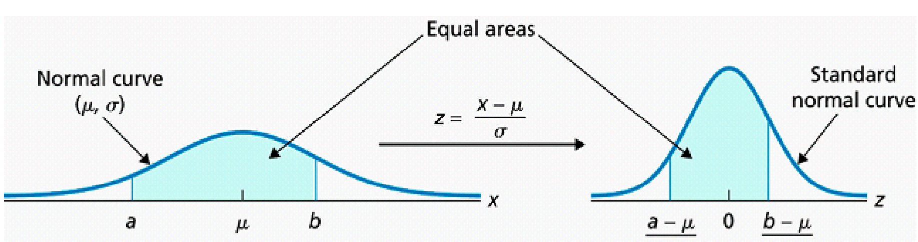
$\mathbf{X} \sim N(\mu\ ,\ \sigma^2) \rightarrow \quad z = \frac{\mathbf{X} - \mu}{\sigma} \sim N(0,1)$

$$\mathbf{P}(\mathrm{X} \le x) = \mathbf{P}\left(\frac{\mathrm{X}-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \mathbf{P}\left(Z \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right) = \mathrm{CDF}\,(\mathbf{Z})$$

$\mathbf{\Phi(z)}$ הינה פונקציית ההסתברות המצטברת (CDF) של ההתפלגות הנורמלית הסטנדרטית $N(0,1)$:

$\mathbf{\Phi(z)} = \mathrm{Prob}\ (\mathbf{Z} \le \mathbf{z})$ עבור $\mathbf{X} \sim N(0,1)$

$$\mathrm{P}(a \le \mathrm{X} \le b) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \le \frac{\mathrm{X}-\mu}{\sigma} \le \frac{b-\mu}{\sigma}\right) = \mathrm{P}\left(\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right) = \mathrm{P}\left(Z \le \frac{b-\mu}{\sigma}\right) - \mathrm{P}\left(Z \le \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$



התפלגות נורמלית סטנדרטית $\mathbf{Z} \sim N(0\,,\,1)$

$\mathbf{\Phi(z)}$ הינה פונקציית ההסתברות המצטברת (CDF) של ההתפלגות הנורמלית הסטנדרטית:

$\Phi(-z) = 1 - \Phi(z)$

$P(-a \leq Z \leq a) = 2\,\Phi(a) - 1$

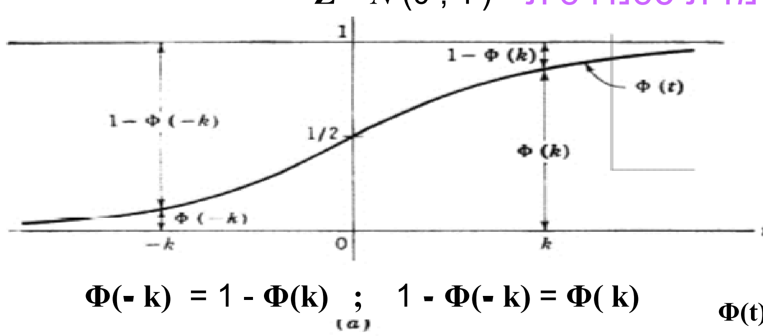
$\Phi(z) = \text{Prob}\,(Z \leq z)$

עבור $Z \sim N(0,1)$

Normal Deviate z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-4.0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.9	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.8	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.7	.0001	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
-3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
-3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483

התפלגות נורמלית סטנדרטית $\mathbf{Z} \sim N(0\,,\,1)$

$\mathbf{\Phi(t)}$ הינה פונקציית ההסתברות המצטברת (CDF) של ההתפלגות הנורמלית הסטנדרטית:

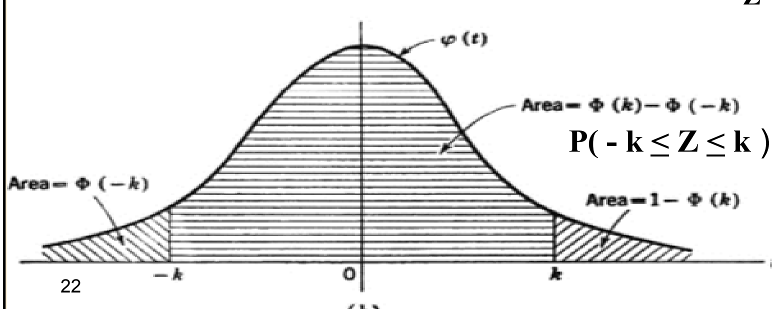


$\Phi(-k) = 1 - \Phi(k)\quad;\quad 1 - \Phi(-k) = \Phi(k)$

(a)

$\Phi(t) = \text{Prob}\,(Z \leq t)$

עבור $Z \sim N(0,1)$



$P(-k \leq Z \leq k) = 2\,\Phi(k) - 1$

22

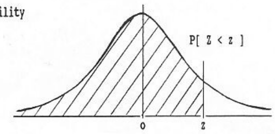
(b)

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\ dZ$$



Z +0.01

Z=1.50

Z=1.51

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981

מה ההסתברות ,
תחת ההתפלגות
הנורמלית הסטנדרטית
עד ל- Z = 1.51 ?
תשובה: 93.45%

מה ההסתברות ,
תחת ההתפלגות
הנורמלית הסטנדרטית
מעל ל- Z = 1.51 ?
תשובה:
= 100% - 93.45%
= 6.55%

מה ההסתברות ,
תחת ההתפלגות
הנורמלית הסטנדרטית
בין Z = 1.51
ל- Z = - 1.51 ?
תשובה:

$= 2\,\Phi(z) - 1$

$= 2 \times 93.45\% - 100\% =$

$= 186.90\% - 100\% =$

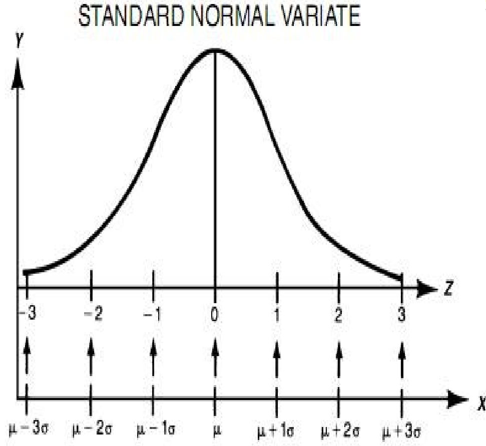
$= 86.90\%$

התפלגות נורמלית סטנדרטית

$\mathbf{Z} \sim N(0\ ,\ 1)$

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

STANDARD NORMAL VARIATE



The Translation of X to Z by the Transformation $Z = (X - \mu)/\sigma$

Figure 3

Standard Normal Probability Content

z	$\int_{-\infty}^{z}$	$\int_{-z}^{z}$	$\int_{z}^{\infty}$
0.0	0.50000	0.00000	0.50000
0.5	0.69146	0.38292	0.30854
1.0	0.84134	0.68269	0.15866
1.5	0.93319	0.86639	0.06681
2.0	0.97725	0.95450	0.02275
2.5	0.99379	0.98758	$6.210 \cdot 10^{-3}$
3.0	0.99865	0.99730	$1.350 \cdot 10^{-3}$
3.5	0.99977	0.99953	$2.326 \cdot 10^{-4}$
4.0	0.99997	0.99994	$3.167 \cdot 10^{-5}$
4.5	1.00000	0.99999	$3.398 \cdot 10^{-6}$
5.0	1.00000	1.00000	$2.867 \cdot 10^{-7}$
6.0	1.00000	1.00000	$9.866 \cdot 10^{-10}$
7.0	1.00000	1.00000	$1.280 \cdot 10^{-12}$
8.0	1.00000	1.00000	$6.221 \cdot 10^{-16}$

z	$\int_{-\infty}^{z}$	$\int_{-z}^{z}$	$\int_{z}^{\infty}$
0.00000	0.5	0.0	0.5
0.25335	0.6	0.2	0.4
0.67449	0.75	0.5	0.25
0.84162	0.8	0.6	0.2
1.28155	0.9	0.8	0.1
1.64485	0.95	0.9	0.05
1.95996	0.975	0.95	0.025
2.32635	0.99	0.98	0.01
2.57583	0.995	0.99	0.005
3.09023	0.999	0.998	0.001
3.29053	0.9995	0.999	0.0005
3.71902	0.9999	0.9998	0.0001
3.89059	0.99995	0.9999	0.00005
4.26489	0.99999	0.99998	0.00001

Probability content for a standard normal distribution for half-integer and integer values of z (left-hand table) and for fixed probability content (right-hand table)

תכונות מועילות ושימושיות נוספות למ"מ בעל התפלגות הנורמלית :

סכום/הפרש שני מ"מ-ים נורמליים ב"ת הינו מ"מ נורמלי בעל תוחלת השווה לסכום/הפרש התוחלות ושונות השווה לסכום השונויות:

$$\mathbf{X_1} \sim N(\mu_1 , \sigma_1^2) \ , \ \mathbf{X_2} \sim N(\mu_2 , \sigma_2^2) \ \rightarrow \ (\mathbf{X_1} + \mathbf{X_2}) \sim N(\mu_1 + \mu_2 , \sigma_1^2 + \sigma_2^2)$$

$$\mathbf{X_1} \sim N(\mu_1 , \sigma_1^2) \ , \ \mathbf{X_2} \sim N(\mu_2 , \sigma_2^2) \ \rightarrow \ (\mathbf{X_1} - \mathbf{X_2}) \sim N(\mu_1 - \mu_2 , \sigma_1^2 + \sigma_2^2)$$

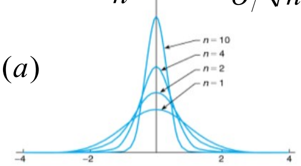
וכללית, קומבינציה לינארית של n מ"מ-ים נורמליים ב"ת, הינה מ"מ נורמלי
בעל תוחלת השווה לקומבינציה הלינארית של תוחלות n המ"מ-ים המרכיבים את הקומבינציה,
ושונות השווה לסכום n מכפלות – כ"א מכפלה של שונות ה-מ"מ בריבוע המקדם שלו בקומבינציה:

עבור $\mathbf{i = 1, 2, \ldots , n}$ $\mathbf{X_i} \sim N(\mu_i , \sigma_i^2)$ ב"ת $\Leftarrow$ $\sum_{i=1}^{n} a_i X_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i \ , \ \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$

לתכונות אילו חשיבות במציאת (הסתברויות) רמות בטחון / סמך ומובהקות באמידה ובדיקת השערות.
למשל במציאת הרווח בר-סמך לאומד לתוחלת (μ) מ"מ נורמלי (כאן, כששונותו ידועה):

$$X_i \sim \text{iid } N(\mu, \sigma^2) \Rightarrow \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}) \Rightarrow \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\Rightarrow P\left(a \le \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le b\right) = \Phi(b) - \Phi(a)$$



Densities of sample means from a standard normal population.

התפלגויות הנגזרות מההתפלגות הנורמלית: χ^2 , t , F

בסטטיסטיקה נדרש להשתמש בהתפלגויות דגימה של אומדי תוחלת ושונות, שלרב הינן התפלגויות הנגזרות מההתפלגות הנורמלית:

התפלגות חי בריבוע - χ^2

ההתפלגות של ריבוע מ"מ נורמלי סטנדרטי (תוחלת 0, שונות 1) הינה התפלגות חי-בריבוע עם דרגת חופש 1.

$$Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2_{(1)}$$

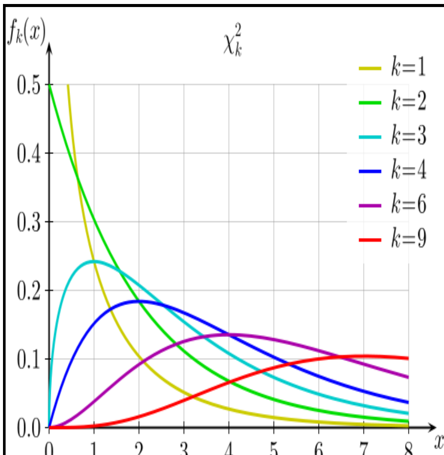
להתפלגות חי-בריבוע תכונת חיבוריות (אדיטיביות) – התפלגות סכום מ"מ-ים ב"ת, כ"א מתפלג חי-בריבוע, הינה התפלגות חי-בריבוע עם סכום דרגות החופש.

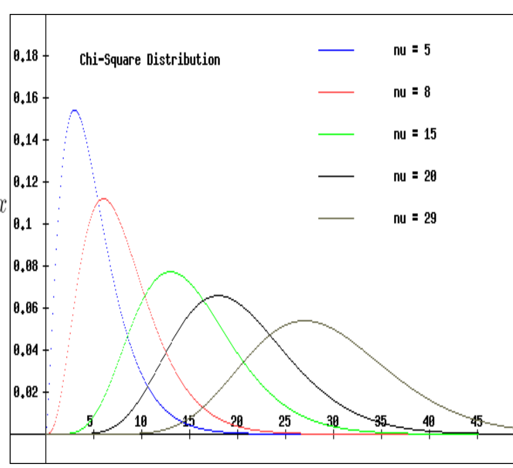
$$Y_i \sim \chi^2_{(n_i)} \text{ indep} \Rightarrow \sum_{i=1}^{m} Y_i \sim \chi^2_{\left(\sum_{i=1}^{m} n_i\right)}$$

לכן, ההתפלגות של סכום ריבועי מ"מ-ים ב"ת נורמליים מתוקננים הינה התפלגות חי-בריבוע עם מספר דרגת חופש כממספר המ"מ-ים הב"ת בסכום.

$$X_i \sim N(\mu,\sigma^2) \text{ indep.} \Rightarrow \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2_{(n)}$$

גרפי התפלגויות חי בריבוע - χ^2





$$E\left(\chi^2_{(n)}\right) = n$$

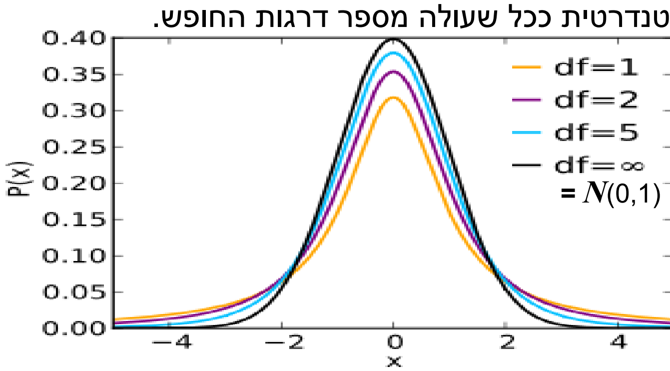
$$\mathrm{Var}\left(\chi^2_{(n)}\right) = 2n$$

התפלגות "סטודנט" 'טי' (מרכזית)' - t "Student" (Central)

$$Z \sim \mathcal{N}(0,1) \quad \text{indep of} \quad Y \sim \chi^2_{(n)}$$

$$\mathbf{t}_{(n)} = \frac{Z}{\sqrt{\chi^2_{(n)}/n}}$$

ההתפלגות של מנת מ"מ-ים ב"ת, כשבמונה מ"מ נורמלי סטנדרטי ובמכנה שורש ריבועי של מ"מ חי-בריבוע המחולק במספר דרגות החופש שלו.
בעלת פרמטר הנקרא **דרגות חופש** השווה לדרגות החופש של המ"מ במכנה.
סימטרית ודומה להתפלגות הנורמלית הסטנדרטית רק עם זנבות עבים יותר.
מתקרבת לנורמלית סטנדרטית ככל שעולה מספר דרגות החופש.



$$E(\mathbf{t}_{(n)}) = 0$$

$$\mathrm{Var}(\mathbf{t}_{(n)}) = \frac{n}{n-2}$$

(המשך) התפלגות 'טי' (מרכזית)' - t Student (Central)

התפלגות t מהווה **ההתפלגות של ממוצע** תצפיות מ-מ"מ נורמלי המתוקנן ע"י תוחלתו μ ואומד בלתי מוטה של סטיית התקן של הממוצע (כלומר, אומד בלתי מוטה של טעות התקן).

$$X_i \sim \mathcal{N}(\mu,\sigma^2) \text{ indep.} \Rightarrow \quad T = \frac{\bar{X}_n - \mu}{\sqrt{S^2/n}} = \frac{\bar{X}_n - \mu}{\sqrt{\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)/n}} =$$

$$= \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\left(\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2\right)/(n-1)}} = \frac{Z}{\sqrt{\sum_{i=1}^{n-1} Z_i^2/(n-1)}} = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{(n-1)}/(n-1)}} \sim \mathbf{t}_{(n-1)}$$

<u>לתשומת לב</u>:
דרגות החופש של ההתפלגות כאן הינה (n-1)
כלומר באחד פחות מגודל המדגם (ממספר התצפיות) – איבדנו דרגת חופש בגלל השימוש בממוצע התצפיות ולא בתוחלת.

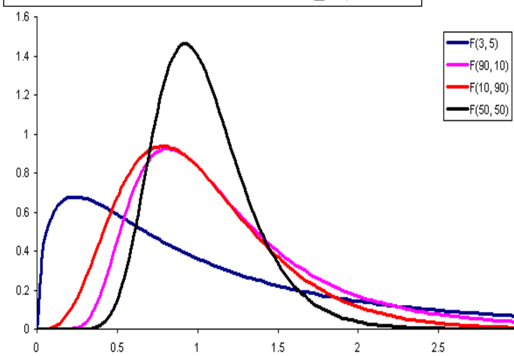
התפלגות (פישר) **סנדקור** F - (Fisher)-**Snedecor's** distribution

$$U \sim \chi^2_{(n_1)} \quad indep.\ of \quad W \sim \chi^2_{(n_2)}$$

$$\mathbf{F}_{(n_1, n_2)} = \frac{U/n_1}{W/n_2} = \frac{\chi^2_{(n_1)} \big/ n_1}{\chi^2_{(n_2)} \big/ n_2}$$

ההתפלגות של מנת שני מ"מ-ים ב"ת, כ"א הינו מ"מ חי-בריבוע המחולק במספר דרגות החופש שלו.

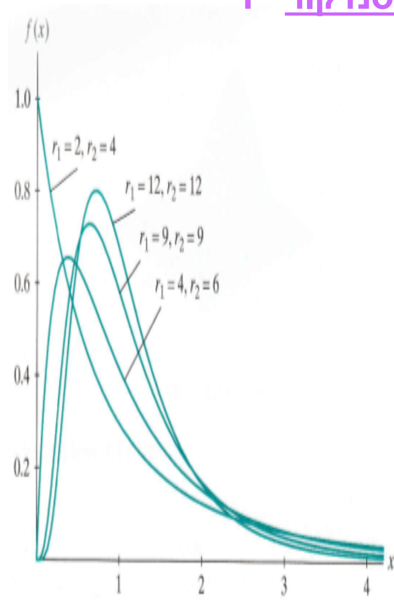
משמש להשוואת אומדני שונויות בבדיקות הומוגניות שונויות ובניתוח שונות (ANOVA)

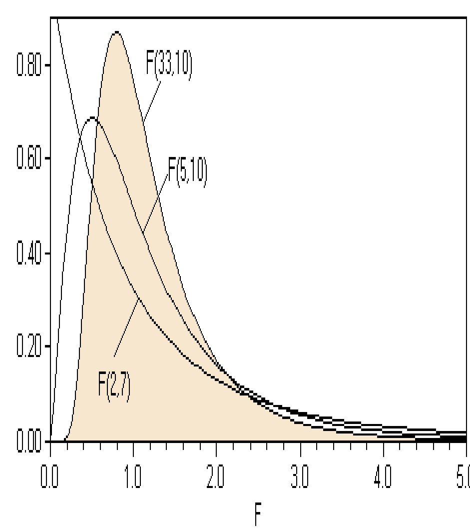


$$E\left(\mathbf{F}_{(n_1, n_2)}\right) = \frac{n_2}{n_2 - 2}$$

$$Var\left(\mathbf{F}_{(n_1, n_2)}\right) = \frac{2 n_2^2 (n_1 + n_2 - 2)}{n_1 (n_2 - 2)^2 (n_2 - 4)}$$

גרפים נוספים של התפלגות (פישר) **סנדקור** F





אי-שיוויון מרקוב - Markov Inequality

$$P(X \geq a) \leq \frac{E(X)}{a}$$

אם $\mathbf{X}$ מ"מ אי-שלילי, אז לכל $a > 0$ מתקיים:

<u>הוכחה</u> יהי $\mathbf{Y}_a$ מ"מ:

$$\mathbf{Y}_a = \begin{cases} 0 & \text{if } \mathbf{X} < a \\ a & \text{if } \mathbf{X} \geq a \end{cases}$$

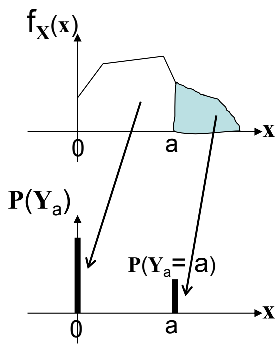
ואז $\mathbf{Y}_a \leq \mathbf{X}$ ולכן גם

$$\mathbf{E}(\mathbf{Y}_a) \leq \mathbf{E}(\mathbf{X})$$

אבל

$$\mathbf{E}(\mathbf{Y}_a) = a\,\mathbf{P}(\mathbf{Y}_a = a) = a\,\mathbf{P}(\mathbf{X} \geq a)$$

ומתקבל: $a\,\mathbf{P}(\mathbf{X} \geq a) \leq \mathbf{E}(\mathbf{X})$ מ.ש.ל.



אי-שיוויון מרקוב אומר שאם למ"מ אי-שלילי
יש ערך תוחלת קטן, ההסתברות לערך גדול גם היא קטנה ויורדת עם עליית גודל הערך.

חסמי אי-שיוויון מרקוב אינם "חדים" - לעיתים, אפילו לא אינפורמטיביים,
למשל כאשר $\mathbf{E}(\mathbf{X}) \geq a$.

תועלת אי-שיוויון מרקוב במקרים בהם התוחלת ידועה, או ניתנת לחישוב בנקל
אך ההתפלגות לא ידועה או קשה לחישוב, למשל במקרה של סכום מ"מ-ים.

אי-שיוויון צ'בישב - Chebyshev's Inequality

$\mathbf{X}$ מ"מ בעל תוחלת $\mathbf{E}(\mathbf{X}) = \mu$ ושונות $\mathbf{Var}(\mathbf{X}) = \sigma^2$, אז לכל $t > 0$

מתקיים: $P(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}$ **או** $P(|X - \mu| < t) \geq 1 - \frac{\sigma^2}{t^2}$

או בניסוח אחר:

$$P(\mu - t\sigma < X < \mu + t\sigma) = P(|X - \mu| < t\sigma) \geq 1 - \frac{1}{t^2}$$

(וזאת רק למאורע סימטרי ביחס לתוחלת). ההוכחה בעזרת אי-שיוויון מרקוב.

אי-שיוויון צ'יבשב אומר שאם למ"מ (כלשהו, לאו דווקא אי-שלילי) שונות קטנה, אז ההסתברות לערך רחוק מהתוחלת גם היא קטנה ויורדת עם עליית המרחק. (כשהמרחק מהתוחלת נמדד במספר סטיות התקן).

חסמי אי-שיוויון צ'יבשב נוטים להיות "חדים" יותר מאילו של אי-שיוויון מרקוב, אך עדיין "גסים" למדי ביחס להסתברויות המדויקות. תועלתו דומה לזו של אי-שיוויון מרקוב (רק כאן צריך גם לדעת את השונות / סטיית התקן).

תועלתו הנוספת הינה היותו הבסיס לחוק המספרים הגדולים (ראו בהמשך).

החוק (החלש) של המספרים הגדולים - The (weak) Law of Large Numbers

יש כמה חוקים (גרסאות) כאילו – כ"א עם תנאים שונים.

חוק זה אומר שבהסתברות גבוהה, ממוצע ערכי (הרבה) תצפיות לא-מתואמות מהתפלגויות בעלות תוחלות ושונויות סופיים ("חסומים"), יהיה קרוב לממוצע התוחלות של התצפיות – ככל שתהיינה יותר תצפיות, ככה ההבדל בין ממוצע ערכי התצפיות וממוצע התוחלות של התצפיות יקטן וההסתברות לכך תגדל.

הגרסה שתשמש אותנו בעיקר נקראת גם **משפט חינצ'ין**:

משפט זה אומר שבהסתברות גבוהה, ממוצע ערכי (הרבה) תצפיות ב"ת בעלי אותה התפלגות ואותה תוחלת ושונות סופיים, יהיה קרוב (מאד) לתוחלת ההתפלגות – ככל שיותר תצפיות, ככה הקירבה תגבר וההסתברות לכך תגדל.

משפט חינצ'ין: תהי $X_1, X_2, X_3, \ldots, X_n, \ldots$ סידרת מ"מ-ים ב"ת המוגדרים על אותו מרחב מדגם, בעלי אותה התפלגות ובעלי אותה תוחלת $E(X_i) = \mu$ אז על הסידרה חל חוק המספרים הגדולים, דהיינו

לכל $\varepsilon > 0$ קיים

$$\text{Prob}\left(\left|\frac{X_1+X_2+X_3+\ldots+X_n}{n} - \mu\right| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1$$

למשפט זה משמעות מיוחדת כשמדובר בניסיונות ברנולי – ראו בהמשך

משפט ברנולי

- מקרה פרטי של משפט חינצ'ין – כאשר התפלגות ה- X_i הינה התפלגות ברנולי. ואז התוחלת $E(X_i) = p$ וכן

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\text{Number of "Successes" in the Sample}}{\text{Number of Observations (Sample Size)}} = \begin{array}{c}\text{Relative frequency}\\ \text{of "Succeses"}\\ \text{in the Sample}\end{array}$$

כלומר, אם נסמן את השכיחות היחסית של ה"הצלחות" במדגם (= ממוצע ערכי התצפיות במדגם) ב- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p}$ נקבל את משפט ברנולי:

משפט ברנולי: תהי $X_1, X_2, X_3, \ldots, X_n, \ldots$ סידרת מ"מ-ים ברנולי,ב"ת

אז על הסידרה חל חוק המספרים הגדולים, דהיינו

לכל $\varepsilon > 0$ קיים:

$$\text{Prob}\left(\left|\hat{p} - p\right| < \varepsilon\right) \xrightarrow[n\to\infty]{} 1$$

וזה, במובן מסוים, מהווה את ההצדקה לקשר (זיהוי ?) בין מושג ההסתברות ומושג השכיחות היחסית (הגישה השכיחותנית - Frequentalist להסתברות).

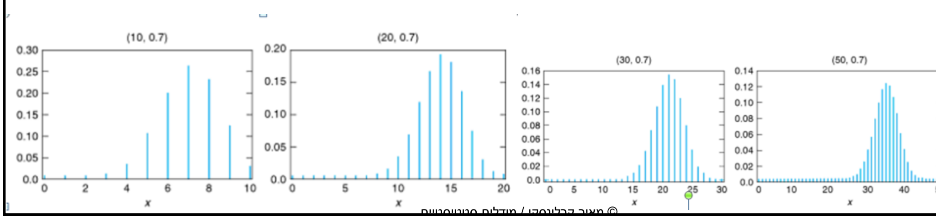
נראה גם שלא רק שהשכיחות היחסית במדגם (שהיא מ"מ) קרובה ל- p ככל שגודל המדגם גדל אלא שהתפלגותה מתכנסת להתפלגות הנורמלית.

משפטי דה-מואבר – De-Moiver ולפלס – Laplace

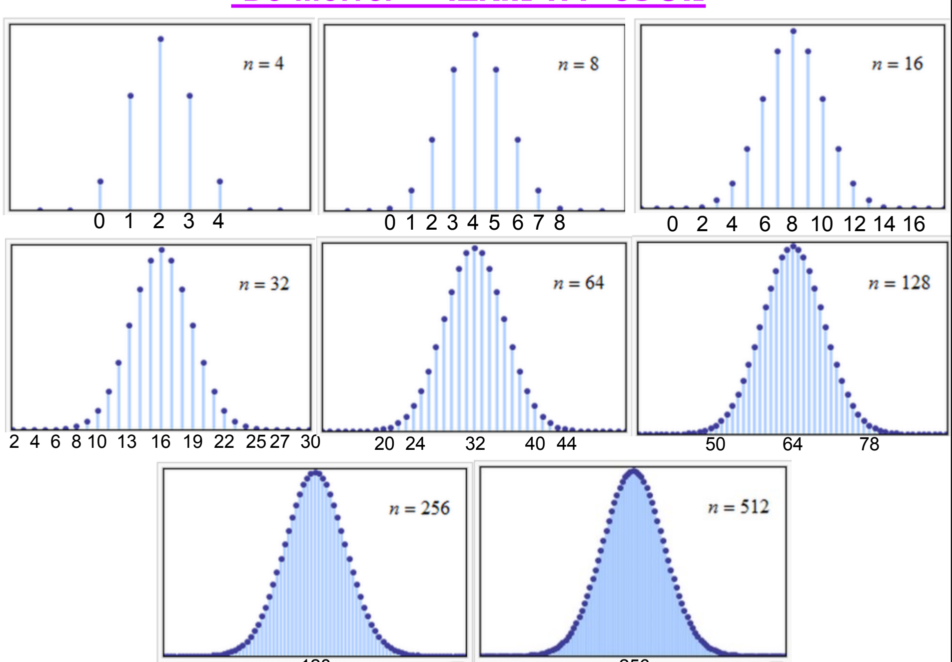
הם משפטים מוקדמים (ומקרים פרטיים) של "משפט הגבול המרכזי" האומרים שככל שמספר התצפיות גדל כן מתקרבת ("מתכנסת") התפלגות המ"מ הבינומי המתוקנן להתפלגות הנורמלית הסטנדרטית:

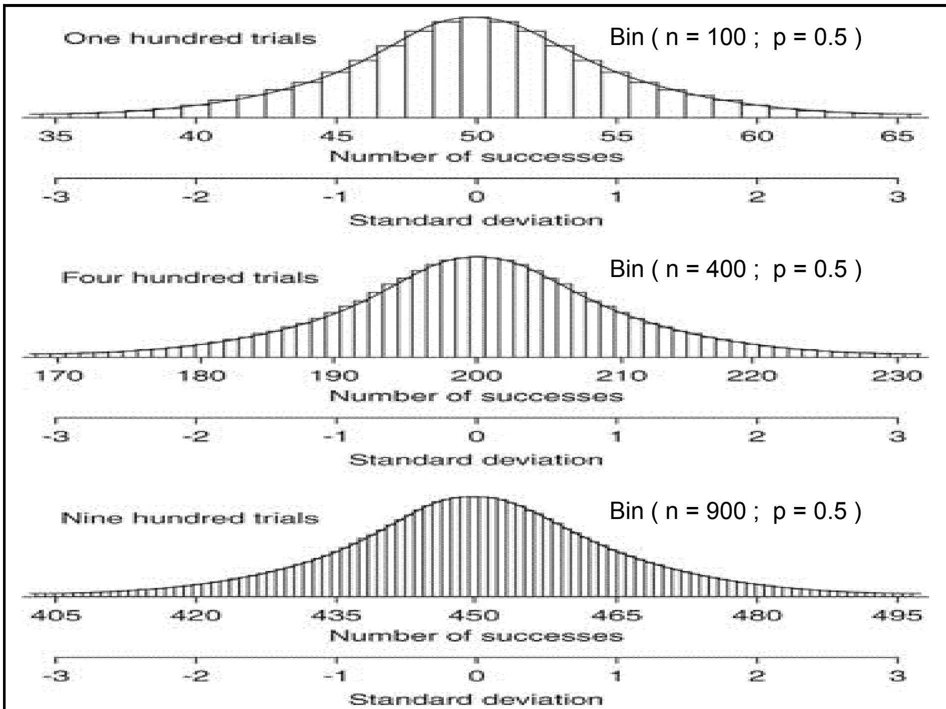
$$\lim_{n\to\infty} P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le a\right) = \Phi(a)$$

$$\lim_{n\to\infty} P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \le a\right) = \Phi(a)$$

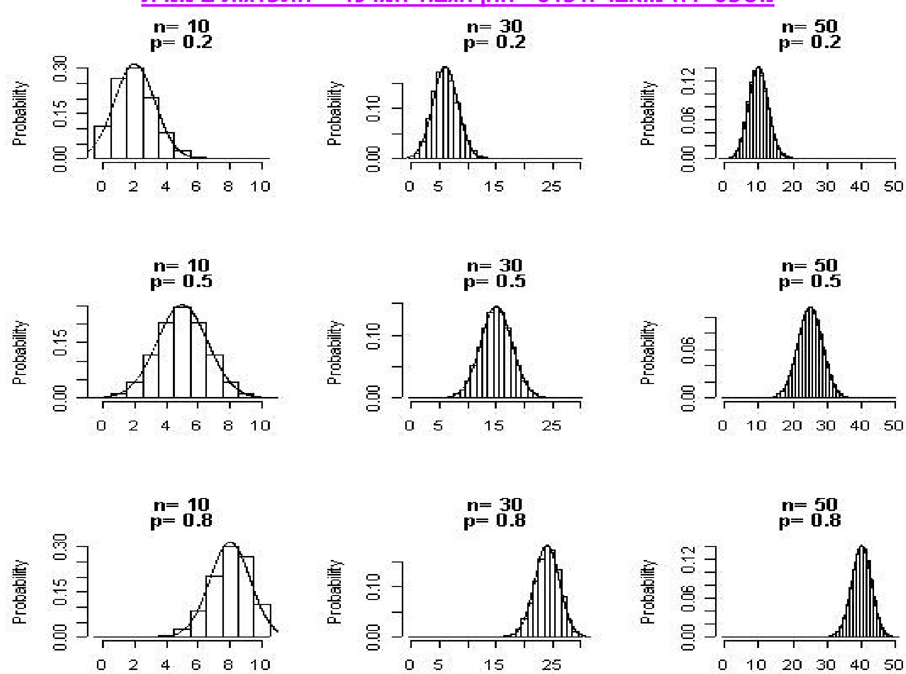


משפט דה-מואבר – De-Moiver





משפטי דה-מואבר ולפלס - חוק הגבול המרכזי – התפלגות בינומית



קירוב נורמלי לבינומי

טיב הקירוב תלוי ב- **p** וב- **n**

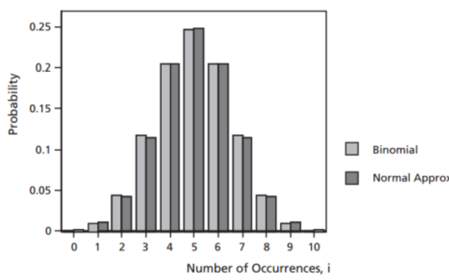


Figure 7.17(a): Comparison at *n* = 10 and *p* = 0.5

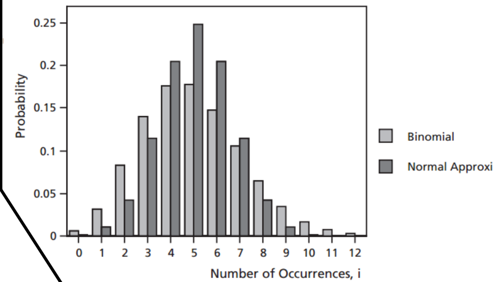


Figure 7.17(c): Comparison at *n* = 250 and *p* = 0.02

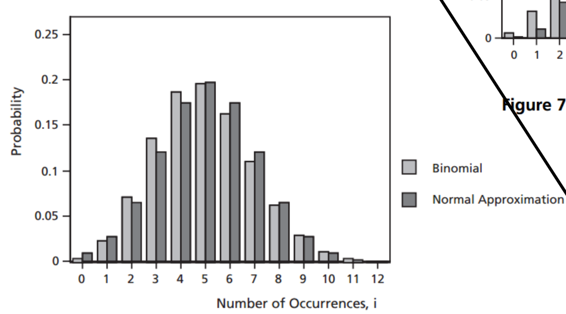


Figure 7.17(b): Comparison at *n* = 25 and *p* = 0.2

קירוב נורמלי לבינומי

במבחן טעם (עיוור), שהשווה בין קוקה-קולה לפפסי-קולה, מתוך 100 נבדקים 56 העדיפו קוקה-קולה ו- 44 העדיפו פפסי-קולה. יהי **Y =** מס' המעדיפים פפסי-קולה.

מטבלאות (או ממחשבון) ההתפלגות הבינומית, עבור $Y \sim Bin(100 ; 0.5)$

$Prob(Y \le 44) = 0.136$

$$Z = \frac{Y - np}{\sqrt{n \times p \times (1-p)}} = \frac{Y/n - p}{\sqrt{\frac{p \times (1-p)}{n}}} = \frac{p_{observed} - p}{\sqrt{\frac{p \times (1-p)}{n}}} = \frac{0.44 - 0.5}{\sqrt{\frac{0.5 \times (1-0.5)}{100}}} =$$

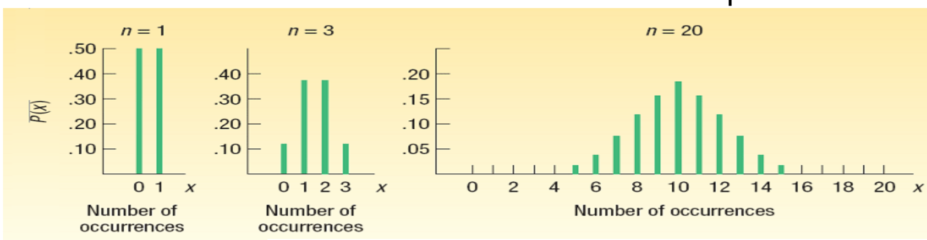
$$= \frac{-0.06}{\sqrt{0.0025}} = \frac{-0.06}{0.05} = -1.2$$

$Z \sim \mathcal{N}(0;1) \Rightarrow Prob(Z \le -1.2) = 0.1151$

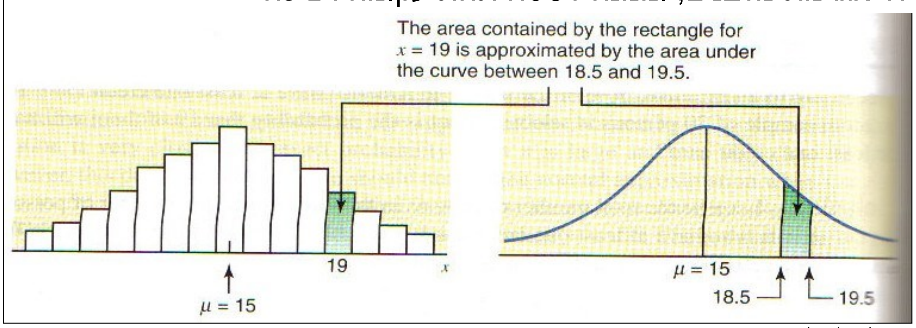
ההסתברויות, בחישוב המדויק (לפי ההתפלגות הבינומית) ובקירוב הנורמלי, די שונות – למה ?

הקירוב הנורמלי לבינומי: תיקון רציפות - Continuity Correction

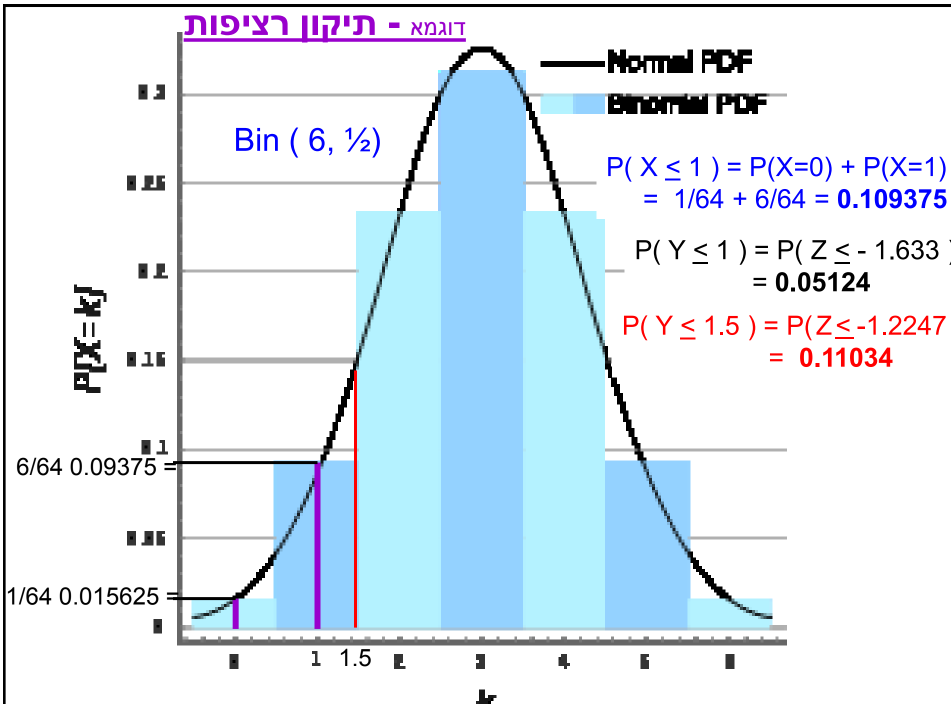
מדיאגרמת "מקלות"



לדיאגרמת מלבנים, וממנה לשטח תחת עקומה רציפה



דוגמא - תיקון רציפות

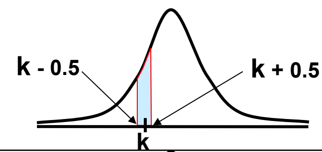
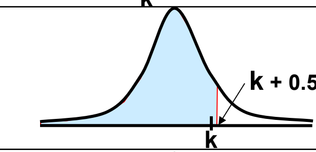
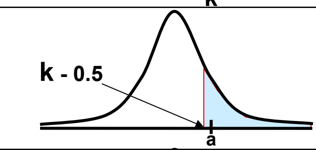
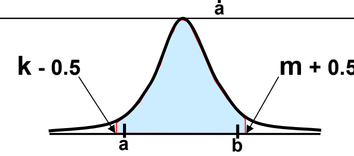


P(X ≤ 1) = P(X=0) + P(X=1)
= 1/64 + 6/64 = **0.109375**

P(Y ≤ 1) = P(Z ≤ - 1.633)
= **0.05124**

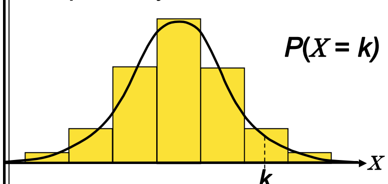
P(Y ≤ 1.5) = P(Z ≤ -1.2247)
= **0.11034**

Continuity Correction - תיקון רציפות

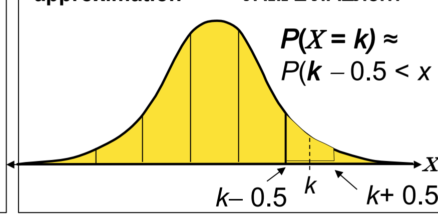
הסתברות בינומית מדויקת (Exact)	קירוב נורמלי להסתברות הבינומית	תיאור גרפי
$P(X = k)$	$P(k-0.5 < X < k+0.5)$	
$P(X \le k)$	$P(X < k+0.5)$	
$P(X \ge k)$	$P(X > k-0.5)$ או $1 - P(X < k-0.5)$	
$P(k \le X \le m)$	$P(k-0.5 < X < m+0.5)$	

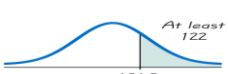
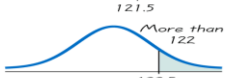
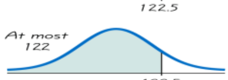
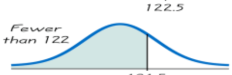
Continuity Correction - תיקון רציפות

Exact binomial probability — הסתברות בינומית מדויקת



Normal approximation — קירוב נורמלי להסתברות בינומית



	הסתברות של	בינומי	קירוב נורמלי
	לפחות k	$P(X \ge k)$	$P(X > k - 0.5)$
	יותר מ- k	$P(X > k)$	$P(X > k + 0.5)$
	לכל היותר k	$P(X \le k)$	$P(X < k + 0.5)$
	פחות מ- k	$P(X < k)$	$P(X < k - 0.5)$

קירוב נורמלי לבינומי (המשך דוגמת החישוב) – **תיקון הרציפות**

אם נוסיף לחישוב בקירוב הנורמלי את "תיקון הרציפות" $Y + ½$
נקבל:

$$\frac{Y+1/2\ -\ np}{\sqrt{np(1-p)}} = \frac{Y/n\ -p+\frac{1}{2*n}}{\sqrt{\frac{p(1-p)}{n}}} = \frac{p_{observed}\ -p+\frac{1}{2*n}}{\sqrt{\frac{p(1-p)}{n}}} =$$

$$= \frac{0.44-0.5+\frac{1}{2*100}}{\sqrt{\frac{0.5*(1-0.5)}{100}}} = \frac{-0.06+0.005}{\sqrt{0.0025}} = \frac{-0.055}{0.05} = -1.1$$

$Z \sim \mathcal{N}(0\ ;1)$ ⟹ Prob ($Z \leq$ - 1.1) = 0.135666 ≈ 0.136

שהתקבל בחישוב הבינומי (ללא תיקון הרציפות קיבלנו Prob =0.1151).

משפט הגבול המרכזי - Central Limit Theorem (CLT)

משפט (תיאורמה) יסודי(ת) בתורת ההסתברות, העוסק בהתפלגות הגבולית של הסכום או הממוצע המצטבר של סדרת משתנים מקריים. בתנאים מסויימים (רחבים למדי ודי שכיחים במציאות), התפלגות הסכום או הממוצע של סדרת משתנים מקריים מתכנסת להתפלגות נורמלית.

למשפט יש כמה נוסחים, חלקם כלליים למדי (למ"מ-ים שוני תוחלות, שוני שוניות ואף שוני התפלגויות) אך ברוב המקרים (ובספרי הלימוד) משתמשים בנוסח המתייחס לסידרת מ"מ-ים שווי התפלגות וב"ת כמו במקרים של מדגם אקראי מאוכלוסייה אינסופית. זהו הנוסח שישמש אותנו.

המשפט:

נתונה סידרת $\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$,... משתנים מקריים בלתי תלויים שווי התפלגות
בעלי תוחלת μ ושונות סופית σ^2 . אזי לכל a (מספר) ממשי מתקיים:

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \leq a\right) = \Phi(a)$$

$$\lim_{n\to\infty} P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) = \Phi(a)$$

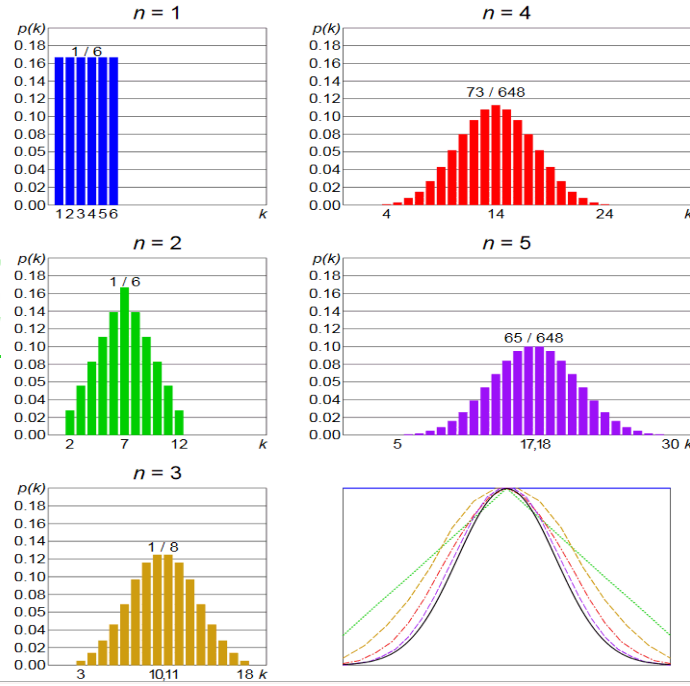
כלומר, סכום המ"מ-ים וממוצעם, מתפלגים כ"א נורמלית עם תוחלות ושוניות בהתאם לקומבינציה שהם.

לא נוכיח משפט זה - הוכחתו (בנוסח הזה) נעשית בעזרת פונקציית יוצרת המומנטים ובד"כ
בעזרת הפונקציה האופיינית – מושגים שהם מעבר לחומר הקורס.
אבל כן נראה גרפית את משמעותו ואת צורתו.

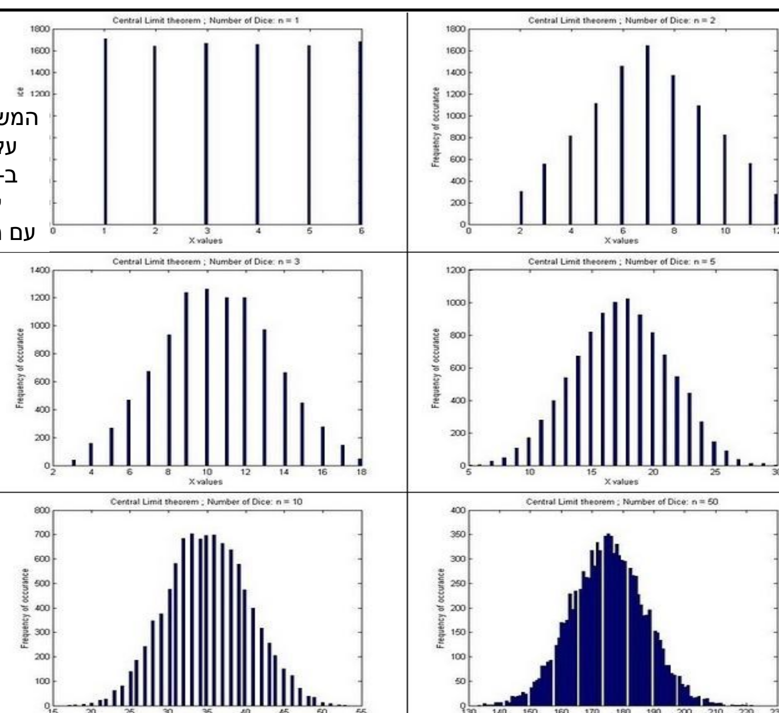
ההתפלגות המקורית
"קובייה הוגנת"

לכל n , מוצגת
התפלגות הדגימה
של סכום הנקודות
במדגמים בגודל n

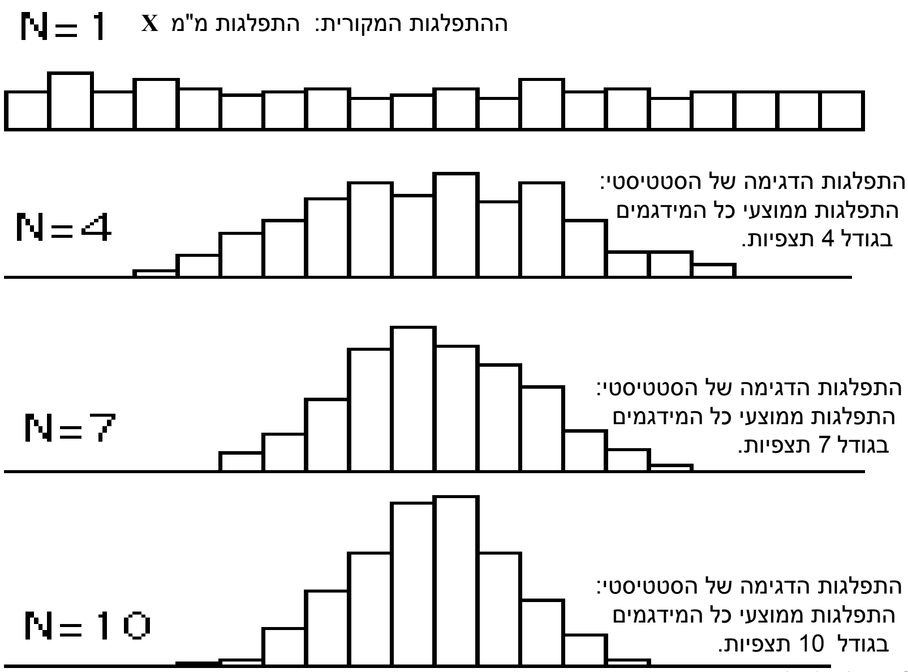
למרות שההתפלגות
המקורית איננה נורמלית
(היא אחידה – בדידה)
נראה בבירור שסכום
תוצאות ההטלות בכל
מדגם (אפילו ל- n=2)
מתפלג בצורה דומה
להתפלגות הנורמלית.
ועוד יותר בשקף הבא -



המשך לשקף הקודם
על סכום נקודות
ב- n הטלות של
קובייה הוגנת
עם מדגם הולך וגדל



הדגמת חוק הגבול המרכזי



Central Limit Theorem (Continued)

The significance of the central limit theorem (C L T) that the distribution of sample means approaches a normal distribution. Refer to Figure 7 below: