

# תרגול 7 - שיערוך פילוג

## בשיטות לא פרמטריות

PDF

Code

### תקציר התיאוריה

#### הגישה הגנרטיבית

בגישה הגנרטיבית אנו נשתמש במדגם על מנת לשערך את הפילוג של  $x$  ו  $y$  מתוך המדגם. על סמך פילוג זה נוכל לבנות חזאי ל  $y$  בהינתן  $x$ .

#### חזאים אידאליים לפונקציות מחיר נפוצות - תזכורת

לרוב אנו נעבוד עם פונקציות מחיר שבהינתן פונקציית הפילוג יש ביטוי סגור לחזאי האידאלי. נזכיר את החזאים האידאליים של פונקציות המחיר הנפוצות:

• **MSE**: התוחלת המותנית:

$$h^*(x) = \mathbb{E}[y|x]$$

• **MAE**: החציון של הפילוג המותנה:

$$h^*(x) = y_{\text{median}} \quad \text{s.t.} \quad F_{y|x}(y_{\text{median}}|x) = 0.5$$

(כאשר  $F_{y|x}$  היא פונקציית הפילוג המצרפי של  $y$  בהינתן  $x$ ).

• **Misclassification rate**: הערך הכי סביר (ה mode):

$$h^*(x) = \arg \max_y p_{y|x}(y|x)$$

#### שימוש בהסתברות המותנית

בעיות סיווג (שבהם  $y$  מקבל סט ערכים בדיד) נוה לשערך את הפילוג המשותף של  $x$  ו  $y$  בעזת הפירוק הבא:

$$p_{x,y}(x,y) = p_{x|y}(x|y)p_y(y)$$

על פי פירוק זה ניתן למעשה לחשב את הפילוג המשותף על ידי כך שנשערך בנפרד את כל אחת מהפילוגים הבאים:

- $p_y(y)$  - הפילוג של  $y$  ללא תלות בערכו של  $x$ . שיערוך זה יהיה לרוב פשוט מכיוון ש  $y$  הוא משתנה דיסקרטי (בדיד).
- $p_{x|y}(x|y)$  כאשר גם כאן יהיה לרוב נוח לפצל את השיערוך למספר שיערוכים שונים בעבור כל ערך אפשרי של  $y$ . זאת אומרת  $p_{x|y}(x|1)$ ,  $p_{x|y}(x|2)$ , וכו'. הדרך לעשות זאת היא על ידי פיצול המדגם על פי הערכים של  $y$  ושיערוך הפילוג של  $x$  בנפרד על כל חלק של המדגם.

#### שיערוך של פונקציות פילוג בשיטות א-פרמטריות

נציג מספר שיטות לשיעור של הסתברויות ופונקציות פילוג של משתנה / וקטור אקראי כל שהוא  $\mathbf{x}$  על סמך מדגם כל שהוא  $\mathcal{D} = \{\mathbf{x}^{(i)}\}$ . בתרגול זה נעסוק בשיטות אשר לא עושות שימוש במודל פרמטרי ולכן הם מכונות א-פרמטריות, בשבוע הבא נעסוק בשיטות פרמטריות.

## מדידה אמפירית / משערך הצבה (Empirical Measure)

המדידה האמפירית,  $\hat{p}_{A,\mathcal{D}}$ , הינה שיעורן של ההסתברות,  $Pr(A)$ , להתרחשות המאורע  $A$ :

$$\hat{p}_{A,\mathcal{D}} = \frac{1}{N} \sum_{i=1}^N I\{\mathbf{x}^{(i)} \in A\}$$

לדוגמא, השיעורן של ההסתברות שהנורמה של  $\mathbf{x}$  קטנה מ-3, זאת אומרת  $A = \{\|\mathbf{x}\|_2 < 3\}$ , תהיה:

$$\hat{p}_{\{\|\mathbf{x}\|_2 < 3\},\mathcal{D}} = \frac{1}{N} \sum_{i=1}^N I\{\|\mathbf{x}^{(i)}\|_2 < 3\}$$

למעשה אנו משערכים כי ההסתברות להתרחשות של מאורע שווה למספר הפעמים היחסי שהמאורע מופיע בסט המדידות.

## שיעורן פונקציית ההסתברות PMF (המקרה של משתנה דיסקרטי)

נוכל לשערך את פונקציית ההסתברות (PMF) של משתנה / וקטור אקראי דיסקרטי על ידי שימוש במדידה האמפירית:

$$\hat{p}_{\mathbf{x},\mathcal{D}}(\mathbf{x}) = \hat{p}_{\{\mathbf{x}=\mathbf{x}\},\mathcal{D}} = \frac{1}{N} \sum_{i=1}^N I\{\mathbf{x}^{(i)} = \mathbf{x}\}$$

## ECDF (Empirical Cumulative Distribution Function)

ECDF הינה שיטה לשערך את פונקציית הפילוג המצרפי (CDF):

$$\hat{F}_{\mathbf{x},\mathcal{D}}(\mathbf{x}) = \hat{p}_{\{\mathbf{x}_j \leq \mathbf{x} \forall j\},\mathcal{D}} = \frac{1}{N} \sum_{i=1}^N I\{\mathbf{x}_j^{(i)} \leq \mathbf{x}_j \forall j\}$$

## היסטוגרמה

היסטוגרמה היא שיטה לשערך פונקציית צפיפות ההסתברות (PDF). שיטה זו נפוצה בעיקר לשם ויזואליזציה של הפילוג של משתנים אקראיים סקלריים. השיעורן מתבצע באופן הבא:

1. מחלקים את תחום הערכים ש  $\mathbf{x}$  יכול לקבל ל bins (תאים) לא חופפים אשר מכסים את כל התחום.
2. לכל תא משערכים את ההסתברות של המאורע ש  $\mathbf{x}$  נמצא בתוך התא.
3. הערך של פונקציית הצפיפות בכל תא תהיה ההסתברות המשוערכת להיות בתא חלקי גודל התא.

נרשום זאת בעבור המקרה של משתנה אקראי סקלרי. נסמן ב  $B$  את מספר התאים וב  $l_b$  ו  $r_b$  את הגבול השמאלי והימני בהתאמה של התא ה  $b$ . ההסטוגרמה תהיה נתונה על ידי:

$$\hat{p}_{\mathbf{x},\mathcal{D}}(\mathbf{x}) = \begin{cases} \frac{1}{\text{size of bin } 1} \hat{p}_{\{\mathbf{x} \text{ in bin } 1\},\mathcal{D}} & \mathbf{x} \text{ in bin } 1 \\ \vdots \\ \frac{1}{\text{size of bin } B} \hat{p}_{\{\mathbf{x} \text{ in bin } B\},\mathcal{D}} & \mathbf{x} \text{ in bin } B \end{cases}$$

$$= \begin{cases} \frac{1}{N(r_1 - l_1)} \sum_{i=1}^N I\{l_1 \leq \mathbf{x}^{(i)} < r_1\} & l_1 \leq \mathbf{x} < r_1 \\ \vdots \\ \frac{1}{N(r_B - l_B)} \sum_{i=1}^N I\{l_B \leq \mathbf{x}^{(i)} < r_B\} & l_B \leq \mathbf{x} < r_B \end{cases}$$

הערות:

- בחירת התאים משפיעה באופן משמעותי על תוצאת השערך של ה PDF.
- כלל אצבע: לחלק את טווח הערכים ל- $\sqrt{N}$  תאים בגודל אחיד.

## (Kernel Density Estimation (KDE

KDE הינה שיטה נוספת לשערוך ה PDF. בשיטה זו אנו נבחר פונקציה המכונה **פונקציית גרעין (kernel)** או **Parzan window** מהם נבנה  $N$  פונקציות גרעין מוזזות בעבור כל נקודה מהמדגם. נסמן ב  $\phi(\mathbf{x})$  את פונקציות הגרעין. פונקציית המוזזות לנקודה ה  $\mathbf{x}^{(i)}$  תהיה  $\phi(\mathbf{x} - \mathbf{x}^{(i)})$ . פונקציית הצפיפות המשוערכת תהיה הממוצע של כל הפונקציות המוזזות:

$$\hat{p}_{\mathbf{x},\phi,\mathcal{D}}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \phi(\mathbf{x} - \mathbf{x}^{(i)})$$

**הערה:** תנאי מספיק והכרחי בכדי שנקבל PDF חוקי, הינו שפונקציית הגרעין תהיה בעצמה PDF חוקי. זאת אומרת שהיא חייבת להיות חיוביות ושהאינטגרל עליה יהיה שווה ל 1.

### הוספת פרמטר רוחב

מקובל להוסיף לפונקציות הגרעין פרמטר  $h$  אשר שולט ברוחב שלה באופן הבא:

$$\phi_h(\mathbf{x}) = \frac{1}{h^D} \phi\left(\frac{\mathbf{x}}{h}\right)$$

החלוקה ב  $h^D$  היא על מנת לשמור על הנרמול של הפונקציה. כאשר  $D$  הוא המימד של  $\mathbf{x}$ .

בתוספת פרמטר זה המשערך יהיה:

$$\hat{p}_{\mathbf{x},\phi,h,\mathcal{D}}(\mathbf{x}) = \frac{1}{Nh^D} \sum_{i=1}^N \phi\left(\frac{\mathbf{x} - \mathbf{x}^{(i)}}{h}\right)$$

### פונקציות גרעין נפוצות

שתי הבחירות הנפוצות ביותר לפונקציית הגרעין הינן:

1. חלון מרובע:

$$\phi_h(\mathbf{x}) = \frac{1}{h^D} I\{|x_j| \leq \frac{h}{2} \quad \forall j\}$$

2. גאוסיאן:

$$\phi_\sigma(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma^D} \exp\left(-\frac{\|\mathbf{x}\|_2^2}{2\sigma^2}\right)$$

כלל אצבע לבחירת רוחב הגרעין במקרה הגאואסי הסקלרי הינו  $\sigma = \left(\frac{4 \cdot \text{std}(\mathbf{x})^5}{3N}\right)^{\frac{1}{5}} \approx 1.06 \text{std}(\mathbf{x})N^{-\frac{1}{5}}$ , כאשר  $\text{std}(\mathbf{x})$  הינה הסטיית תקן של  $\mathbf{x}$  (אשר לרוב תהיה משוערכת גם היא מתוך המדגם)

## תוחלת אמפירית (Empirical mean)

התוחלת האמפירית משערכת את התוחלת של פונקציה מסויימת של המשתנה האקראי  $f(\mathbf{x})$ , על ידי החלפת התוחלת במיצוע של הפונקציה על הדגימות במדגם:

$$\hat{\mu}_{f(\mathbf{x}),\mathcal{D}} = \frac{1}{N} \sum_{i=1}^N f(\mathbf{x}^{(i)})$$

## ה bias וה variance של משערך

כפי שצינו כאשר עסקנו ב bias-variance tradeoff, בכדי לשערך את הביצועים של שיטה מסויימת נרצה להסתכל על הפילוג של תוצאות השערך הנובע מהאקראיות של המדגם. נשתמש שוב בסימון  $\mathbb{E}_{\mathcal{D}}$  בכדי לסמן תוחלת על פני הפילוג של המדגם.

## Bias

בעבור שיערוך של גודל כל שהוא  $z$  בעזרת משעריך  $\hat{z}_D$ , ה bias (היסט) של השיערוך מוגדר כ:

$$\text{Bias}(\hat{z}) = \mathbb{E}_D[\hat{z}_D] - z$$

כאשר ההטיה שווה ל-0, אנו אומרים שהמשעריך אינו מוטא (Unbiased).

## Variance

ה variance (שונות) של המשעריך יהיה:

$$\text{Var}(\hat{z}) = \mathbb{E}_D[(\hat{z}_D - \mathbb{E}_D[\hat{z}_D])^2] = \mathbb{E}_D[\hat{z}_D^2] - \mathbb{E}_D[\hat{z}_D]^2$$

אנו נהיה מעוניינים כמובן במשעריך שגם ה bias וגם ה variance שלו קטנים.

## תרגיל 7.1 - משתנה בינארי (ברנולי)

**(1)** המשתנה האקראי  $x$  הוא משתנה בינארי (משתנה אשר יכול לקבל את הערכים 0 או 1). נתון לנו מדגם המכיל  $N$  דגימות של  $x$ . חשבו את השיערוך של פונקציית ההסתברות של  $x$ . בטאו את התשובה בעזרת  $N_0$  ו  $N_1$ , כאשר  $N_0$  הוא מספר הדגימות ששוות ל 0 ו  $N_1$  הוא מספר הדגימות ששוות ל 1.

נתון כי הפילוג האמיתי של  $x$  הינו:

$$p_x(x) = \begin{cases} 1 & p \\ 0 & (1-p) \end{cases}$$

שני הסעיפים הבאים לא קשורים למדגם הנתון.

**(2)** חשבו את ה bias של המשעריך ב  $x = 1$ .

**(3)** חשבו את ה variance של המשעריך ב  $x = 1$ .

## פתרון 7.1

**(1)**

השיערוך של פונקציית ההסתברות בעבור  $x = 0$  הינו:

$$\hat{p}_{x,D}(0) = \frac{1}{N} \sum_{i=1}^N I\{x^{(i)} = 0\} = \frac{N_0}{N}$$

ובאופן דומה

$$\hat{p}_{x,D}(1) = \frac{1}{N} \sum_{i=1}^N I\{x^{(i)} = 1\} = \frac{N_1}{N}$$

סה"כ

$$p_x(x) = \begin{cases} \frac{N_1}{N} & x = 1 \\ \frac{N_0}{N} & x = 0 \end{cases}$$

**(2)**

נחשב את התחולת של המשעריך  $\hat{p}_{x,D}(1)$ :

$$\mathbb{E}_{\mathcal{D}} [\hat{p}_{x,\mathcal{D}}(1)] = \mathbb{E}_{\mathcal{D}} \left[ \frac{1}{N} \sum_{i=1}^N I\{x^{(i)} = 1\} \right]$$

שימו לב שבחישוב זה אנו לא מתייחסים ל  $x^{(i)}$  כאל מספר ידוע אלא כאל משתנה אקראי. נוציא את החלוקה ב  $N$  ואת הסכימה אל מחוץ לתחלת:

$$= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathcal{D}} [I\{x^{(i)} = 1\}]$$

משום שכל ה  $x^{(i)}$  הם משתנים אקראיים זהים ומפולגים לפי הפילוג של  $x$ , ניתן להסיר את האינדקס של  $(i)$ :

$$\begin{aligned} &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\mathcal{D}} [I\{x = 1\}] \\ &= \mathbb{E}_{\mathcal{D}} [I\{x = 1\}] = p \end{aligned}$$

ה bias יהיה:

$$\text{Bias}(\hat{p}_x(1)) = \mathbb{E}_{\mathcal{D}} [\hat{p}_{x,\mathcal{D}}(1)] - p = p - p = 0$$

מכאן שהמשערך של ההסתברות של משתנים בדידים הוא **משערך לא מוטה**.

### 3

נחשב את התחולת של  $\hat{p}_{x,\mathcal{D}}(1)^2$ :

$$\begin{aligned} \mathbb{E}_{\mathcal{D}} [\hat{p}_{x,\mathcal{D}}(1)^2] &= \mathbb{E}_{\mathcal{D}} \left[ \left( \frac{1}{N} \sum_{i=1}^N I\{x^{(i)} = 1\} \right)^2 \right] \\ &= \frac{1}{N^2} \mathbb{E}_{\mathcal{D}} \left[ \left( \sum_{i=1}^N I\{x^{(i)} = 1\} \right) \left( \sum_{j=1}^N I\{x^{(j)} = 1\} \right) \right] \\ &= \frac{1}{N^2} \sum_{i,j} \mathbb{E}_{\mathcal{D}} [I\{x^{(i)} = 1\} I\{x^{(j)} = 1\}] \end{aligned}$$

מכיוון שבעבור  $i \neq j$  המשתנים  $x^{(i)}$  ו  $x^{(j)}$  הם משתנים בלתי תלויים, נוכל במקרים אלו לפרק את התחולת של המכפלה למכפלת התחולות. נפריד אם כן את הסכום למקרים בהם  $i = j$  (יש  $N$  מקרים כאלה) ולמקרים שבהם  $i \neq j$  (יש  $N^2 - N$  מקרים כאלה):

$$\begin{aligned} &= \frac{1}{N^2} \sum_i \mathbb{E}_{\mathcal{D}} [I\{x^{(i)} = 1\} I\{x^{(i)} = 1\}] + \frac{1}{N^2} \sum_{i \neq j} \mathbb{E}_{\mathcal{D}} [I\{x^{(i)} = 1\} I\{x^{(j)} = 1\}] \\ &= \frac{1}{N^2} \sum_i \mathbb{E}_{\mathcal{D}} [I\{x^{(i)} = 1\}] + \frac{1}{N^2} \sum_{i \neq j} \mathbb{E}_{\mathcal{D}} [I\{x^{(i)} = 1\}] \mathbb{E}_{\mathcal{D}} [I\{x^{(j)} = 1\}] \end{aligned}$$

בדומה לסעיף הקודם נוכל להסיר את האינדקסים:

$$\begin{aligned} &= \frac{1}{N} \mathbb{E}_{\mathcal{D}} [I\{x = 1\}] + \frac{N^2 - N}{N^2} \mathbb{E}_{\mathcal{D}} [I\{x = 1\}]^2 \\ &= \frac{1}{N} p + \frac{N^2 - N}{N^2} p^2 \\ &= \frac{1}{N} (p - p^2) + p^2 \\ &= \frac{1}{N} p(1 - p) + p^2 \end{aligned}$$

ה variance יהיה:

$$\text{Var}(\hat{p}_x(1)) = \mathbb{E}_{\mathcal{D}} [\hat{p}_{x,\mathcal{D}}(1)^2] - \mathbb{E}_{\mathcal{D}} [\hat{p}_{x,\mathcal{D}}(1)]^2 = \frac{1}{N}p(1-p) + p^2 - p^2 = \frac{1}{N}p(1-p)$$

כפי שהיינו מצפים ניתן לראות כי השונות הולכת וקטנה עם מספר הדגימות, שכן ככל שיש לנו יותר דגימות כך השיעור יהיה מדויק יותר. בנוסף, בתור אימות, ניתן להבחין כי בעבור  $N = 1$  נקבל שהשיעור הוא הערך של הדגימה היחידה ובמקרה זה השונות בדיוק שווה לשונות של משתנה בינארי  $p(1-p)$ .

## תרגיל 7.2 - EDCF

בעבור משתנה אקראי רציף כל שהוא  $x$ , מהו ה bias וה variance של משעריך ה EDCF בנקודה מסוימת  $x_0$ ? בטאו את התשובה בעזרת הפילוג המצרפי האמיתי

### פתרון 7.2

למעשה לפתרון תרגיל זה נוכל להשתמש בתוצאת הסעיף הקודם. שיעור ה EDCF בנקודה  $x_0$  נתון על ידי:

$$\hat{F}_{x,\mathcal{D}}(x_0) = \hat{p}_{\{x \leq x_0\},\mathcal{D}}$$

נוכל אם כן אז להגדיר משתנה אקראי בינארי חדש  $z$  אשר שווה ל-1 אם  $x \leq x_0$  ו-0 אחרת. בעזרת משתנה זה נוכל לכתוב את שיעור ה EDCF כשיעור של ההסתברות ש  $z = 1$ :

$$\hat{F}_{x,\mathcal{D}}(x_0) = \hat{p}_{\{z=1\},\mathcal{D}} = \hat{p}_{z,\mathcal{D}}(1)$$

את ה bias וה variance של המשעריך הזה חישבנו בסעיף הקודם וקיבלנו ש:

$$\text{Bias}(\hat{p}_z(1)) = 0$$

$$\text{Var}(\hat{p}_z(1)) = \frac{1}{N}p(1-p)$$

כאשר  $p$  הוא ההסתברות האמתית ש  $z = 1$ . במקרה שלנו  $p = F_x(x_0)$ , ולכן נקבל ש:

$$\text{Bias}(\hat{F}_x(x_0)) = 0$$

$$\text{Var}(\hat{F}_x(x_0)) = \frac{1}{N}F_x(x_0)(1 - F_x(x_0))$$

## תרגיל 7.3 - פילוג משותף

נתון כי  $y$  הינו משתנה אקראי בינארי ו  $x$  משתנה אקראי רציף אשר יכול לקבל ערכים בתחום  $[0, 15]$ . כמו כן נתון לנו המדגם הבא של זוגות של  $x$  ו  $y$ :

	1	2	3	4	5	6	7
x	1	7	9	12	4	4	7
y	0	0	0	0	1	1	1

**1** חשבו את הפילוג המשותף של  $x$  ו  $y$  על ידי שימוש בהסטוגרמה לשיעור של  $x$  בהינתן  $y$ . חלקו את התחום  $[0, 15]$  לשלושה חלקים שווים.

**2** בעבור  $x = 6$  מהו החיזוי האופטימאלי של  $y$  תחת פונקציית המחיר של missclassification rate.

**3** חזרו על שני הסעיפים עם הסטוגרמה שמחלקת את התחום ל-15 תאים.

**4** חזרו על שני הסעיפים הראשונים עם KDE עם פונקציית גרעין של מסוג חלון מרובע ופרמטר רחב  $h = 5$

### פתרון 7.3

# (1)

נחשב את הפילוג המשותף על ידי שימוש בתוחלת המותנית:

$$p_{x,y}(x,y) = p_{x|y}(x|y)p_y(y)$$

$p_y$

נתחיל בלשערך את  $p_y$ . מכיוון ש  $y$  הוא משתנה בינארי, השיערוך של הפילוג שלו יהיה:

$$p_y(y) = \begin{cases} \frac{N_1}{N} = \frac{3}{7} & y = 1 \\ \frac{N_0}{N} = \frac{4}{7} & y = 0 \end{cases}$$

השיערוך של  $p_{x|y}(x|y)$  הוא למעשה שני שיערוכים של שתי פונקציות פילוג,  $p_{x|y}(x|0)$  ו  $p_{x|y}(x|1)$ . נתחיל מהמקרה של  $y = 0$

$p_{x|y}(x|0)$

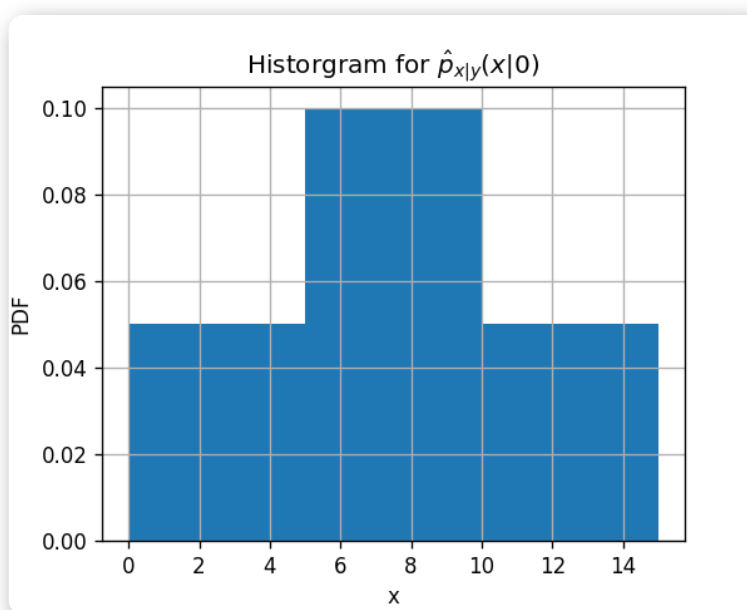
נסתכל רק על הדגימות שבהן  $y^{(i)} = 0$ . ישנם ארבע דגימות כאלה. על פי ההנחה נחלק את התחום ל-3 תאים שווים,  $[0, 5]$ ,  $[5, 10]$  ו  $[10, 15]$ . נחשב את צפיפות ההסתברות בכל תא בעזרת היסטוגרמה. על פי הגדרת היסטוגרמה הצפיפות הסתברות בכל תא שווה לכמות הדגימות מהמדגם ששייכות לתא זה חלקי מספר הדגימות הכולל, חלקי גודל התא.

מתוך הדגימות שבהם  $y^{(i)} = 0$  ישנה דגימה בודד שהגיעה לתא של  $[0, 5]$  ולכן צפיפות ההסתברות בתא זה תהיה:

$$\frac{1}{4(5-0)} = 0.05$$

באופן דומה נחשב את הצפיפות ההסתברות בשאר התאים:

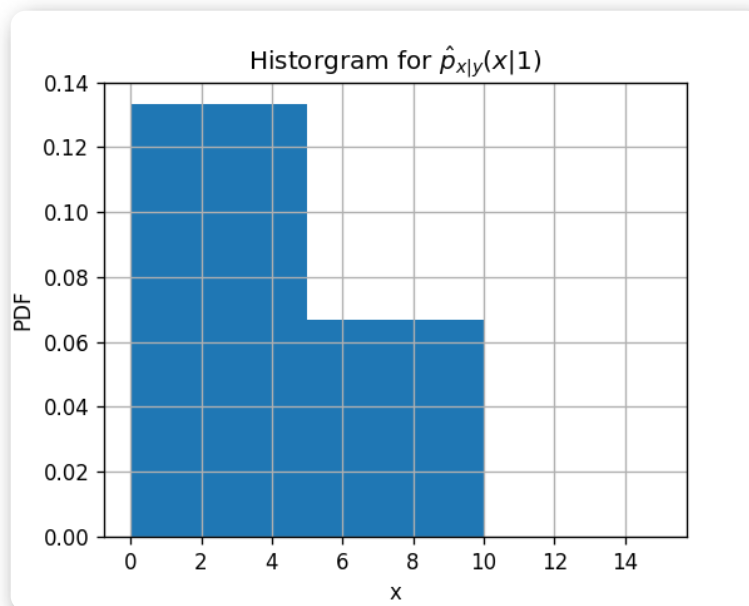
$$\hat{p}_{x|y,D}(x|0) = \begin{cases} \frac{1}{4(5-0)} = \frac{1}{20} & 0 \leq x < 5 \\ \frac{2}{4(10-5)} = \frac{1}{10} & 5 \leq x < 10 \\ \frac{1}{4(15-10)} = \frac{1}{20} & 10 \leq x < 15 \end{cases}$$



$p_{x|y}(x|1)$

בעבור הדגימות שבהם  $y^{(i)} = 1$  נקבל:

$$\hat{p}_{x|y,D}(x|1) = \begin{cases} \frac{2}{3(5-0)} = \frac{2}{15} & 0 \leq x < 5 \\ \frac{1}{3(10-5)} = \frac{1}{15} & 5 \leq x < 10 \\ \frac{0}{3(15-10)} = 0 & 10 \leq x < 15 \end{cases}$$



הפילוג המשותף יהיה אם כן:

	$0 \leq x < 5$	$5 \leq x < 10$	$10 \leq x \leq 15$
$y = 0$	$\frac{1}{20} \frac{4}{7} = \frac{1}{35}$	$\frac{1}{10} \frac{4}{7} = \frac{2}{35}$	$\frac{1}{20} \frac{4}{7} = \frac{1}{35}$
$y = 1$	$\frac{2}{15} \frac{3}{7} = \frac{2}{35}$	$\frac{1}{15} \frac{3}{7} = \frac{1}{35}$	$0 \frac{3}{7} = 0$

**(2)**

אנו יודעים כי החזאי האופטימאלי תחת פונקציית המחיר של misclassification rate הינו הערך הכי סביר של  $y$  בהינתן  $x$ . אם כן עלינו להשוות בין  $p_{y|x}(1|6)$  לבין  $p_{y|x}(0|6)$ .

באופן עקרוני עלינו לחשב את:

$$p_{y|x}(y|x) = \frac{p_{x,y}(x,y)}{p_x(x)} = \frac{p_{x|y}(x|y)p_y(y)}{p_x(x)}$$

אך נשיב לב שהמכנה אינו משנה כלל לתוצאה מפני שהוא משותף לשתי ההסתברויות המותנות שברצונו להשוות ולכן מספיק להסתכל על:

$$p_{y|x}(0|6) \propto p_{x|y}(6|0)p_y(0) = \frac{1}{10} \frac{4}{7} = \frac{2}{35}$$

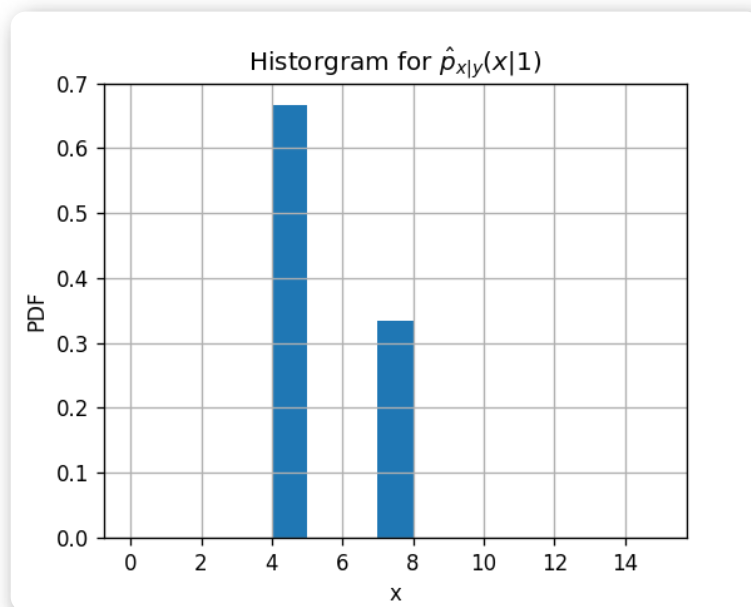
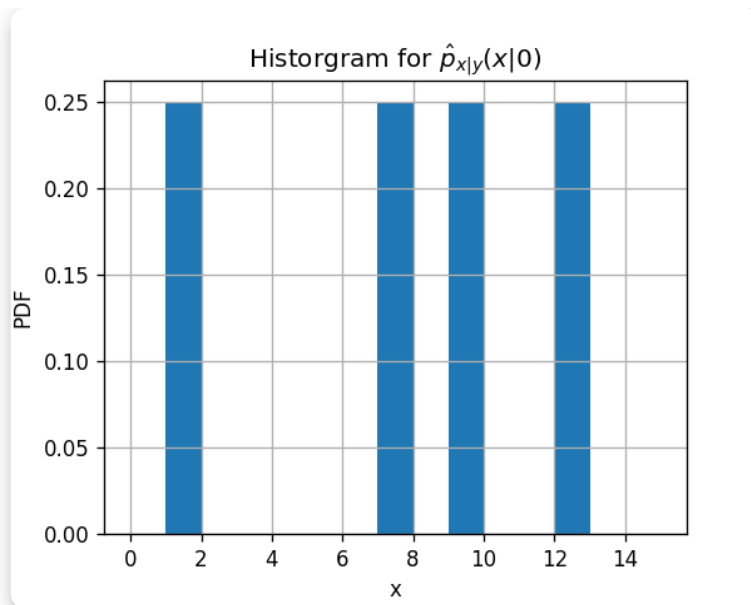
$$p_{y|x}(1|6) \propto p_{x|y}(6|1)p_y(1) = \frac{1}{15} \frac{3}{7} = \frac{1}{35}$$

ולכן הערך היותר סביר הוא 0 וזה יהיה החיזוי שלנו.

**(3)**

נחשב את הפילוג המשותף באופן דומה ונקבל:





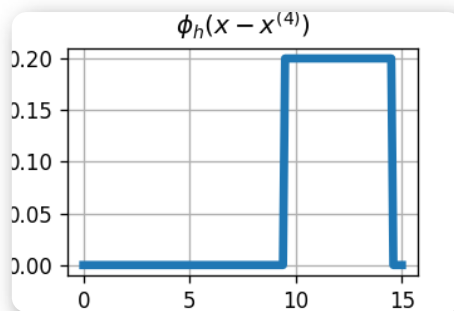
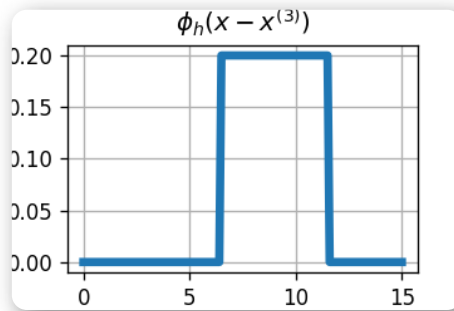
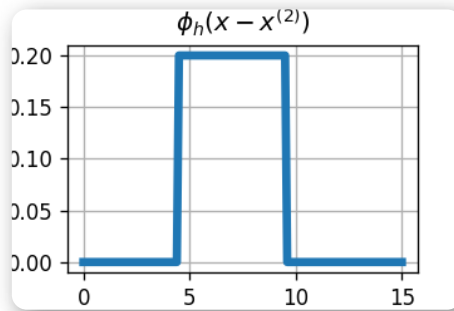
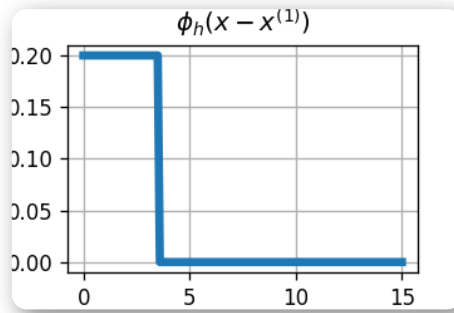
	$0 \leq x < 1$	$1 \leq x < 2$	$2 \leq x < 3$	$3 \leq x < 4$	$4 \leq x < 5$	$5 \leq x < 6$	$6 \leq x < 7$	$7 \leq x < 8$	$8 \leq x < 9$	$9 \leq x < 10$	$10 \leq x < 11$	$11 \leq x < 12$	$12 \leq x < 13$	$13 \leq x < 14$	$14 \leq x < 15$
$y = 0$	0	$\frac{1}{7}$	0	0	0	0	0	$\frac{1}{7}$	0	$\frac{1}{7}$	0	0	$\frac{1}{7}$	0	0
$y = 1$	0	0	0	0	$\frac{2}{7}$	0	0	$\frac{1}{7}$	0	0	0	0	0	0	0

בפילוג זה גם  $p_{y|x}(1|6)$  וגם  $p_{y|x}(0|6)$  שווים ל0 ולכן שני הערכים של y סבירים באותה המידה.

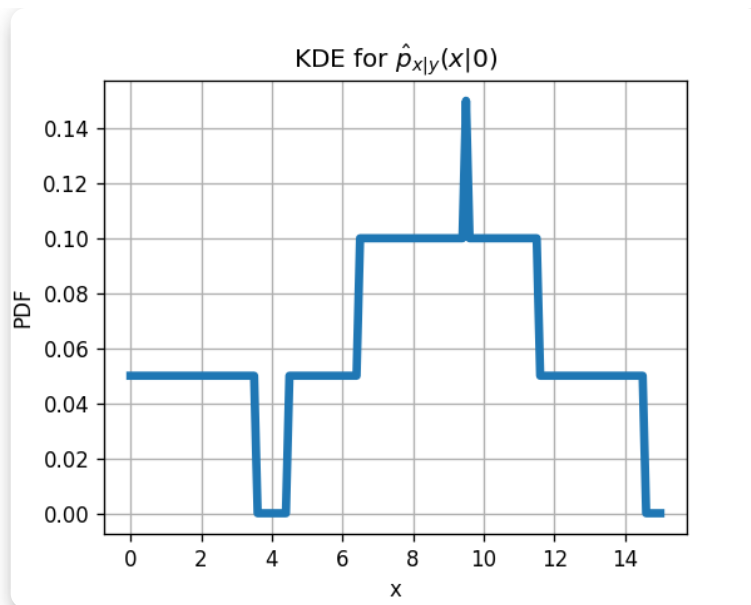
הבעיה עם הפילוג הזה הינה שנראה שלקחנו כמות תאים גדולה מידי ולכן ברוב התאים אין לנו דגימות בכלל וכנראה שהשיעור שם לא מייצג כלל את הפילוג האמיתי.

**(4)**

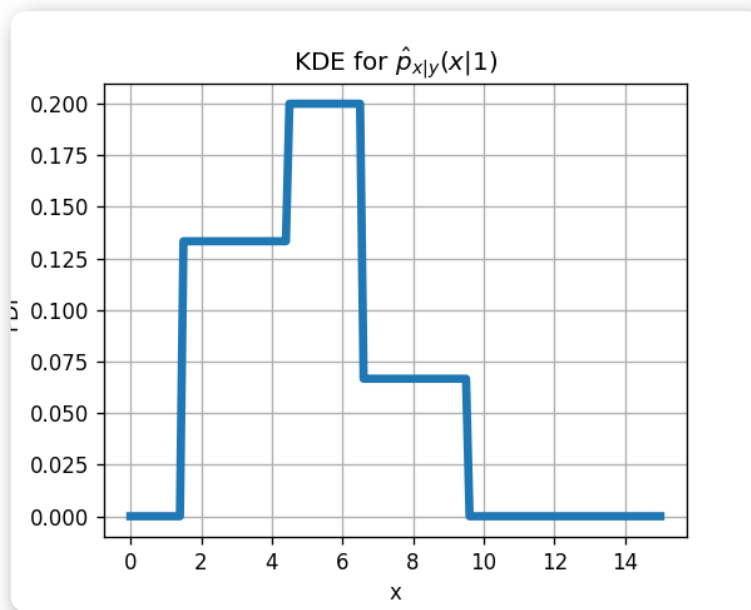
בכדי לבנות כעת את פונקציות הפילוג של  $p_{x|y}(x|y)$  עלינו לקחת כל נקודה מהמדגם (עם ה  $y$  המתאים) ולמקם סביבה חלון ריבועי ברוחב 5 ובגובה  $\frac{1}{5}$ . החלונות של הדגימות המתאימות ל  $y = 0$  הם:



פונקציית הפילוג תהיה הממוצע של כל החלונות הריבועיים:



ובאופן דומה בעבור  $y = 1$ :



מכאן ש:

$$p_{x|y}(6|0) = \frac{1}{20}$$

$$p_{x|y}(6|1) = \frac{1}{5}$$

וההסתברות המותנית של  $x$  ב  $y$  היא

$$p_{y|x}(0|6) \propto p_{x|y}(6|0)p_y(0) = \frac{1}{20} \frac{4}{7} = \frac{1}{35}$$

$$p_{y|x}(1|6) \propto p_{x|y}(6|1)p_y(1) = \frac{1}{5} \frac{3}{7} = \frac{3}{35}$$

לכן הערך הסביר יותר הינו  $y = 1$ .

למעשה בעבור כל שיטת שיערוך קיבלנו תוצאה שונה. עובדה זו מחזקת את הנקודה שלשיערוכים שנקבל ישנה תלות גבוהה בשיטה שנבחר להשתמש בה.

## תרגיל מעשי - שיערוך הפילוג של זמני נסיעה בניו יורק

Code

נחזור למדגם של נסיעות המונית בניו יורק:

ay of ek	duration	dropoff northing	dropoff easting	pickup northing	pickup easting	tip amount	fare amount	payment type	trip distance	passenger count
3	11.5167	4515.18	588.155	4512.98	586.997	0	9.5	2	2.76806	2
6	12.6667	4512.63	584.85	4512.92	587.152	0	10	2	3.21868	1
0	5.51667	4513.17	585.434	4513.36	587.005	2.49	7	1	2.57494	1
1	9.88333	4512.55	586.672	4511.73	586.649	1.65	7.5	1	0.965604	1
2	8.68333	4511.76	585.262	4511.89	586.967	1.66	7.5	1	2.46229	1
3	9.43333	4511.54	585.169	4512.88	585.926	2.2	7.5	1	1.56106	5
5	7.95	4514.21	588.71	4515.08	586.731	1	8	1	2.57494	1
5	4.95	4509.55	585.844	4509.71	585.345	0	5	2	0.80467	1
5	11.0667	4507.74	583.671	4509.48	585.422	1.1	10	1	3.6532	1
3	4.21667	4513.71	587.701	4514.93	587.875	1.36	5.5	1	1.62543	6

בתרגול זה אנו נשתמש רק בשני השדות הבאים:

- duration**: משך הנסיעה הכולל בדקות.
- timeofday**: שעת תחילת הנסיעה כמספר (לא שלם)

(תיאור מלא של כל השדות בטבלה ניתן למצוא פה)

### המשימה: שיערוך הפילוג של זמן הנסיעה של מוניות

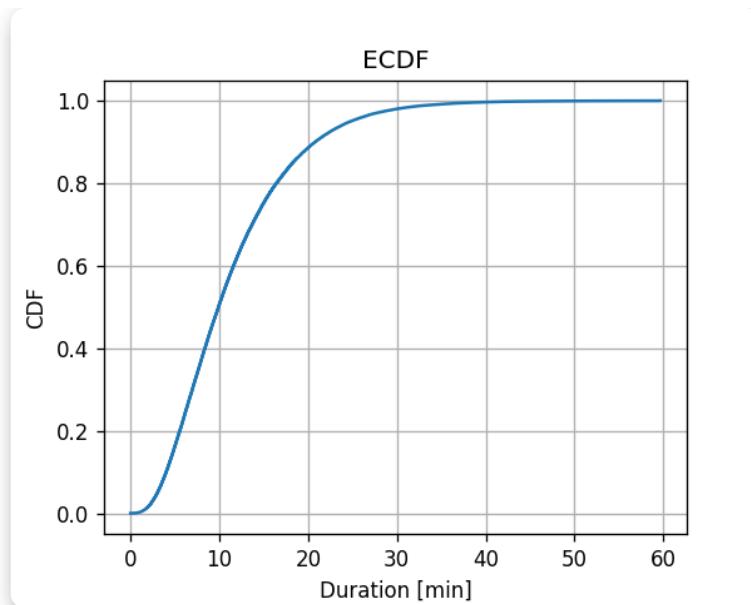
נהג מונית מעוניין לשערך את הפילוג של משך הנסיעות שלו. הוא לקח את הקורס מבוא למערכות לומדות והוא יודע שהוא יוכל לעשות זאת מתוך המידע ההיסטורי אותו אספה עיריית New York. בחלק זה של התרגול אנו נעזור לאותו נהג מונית לבצע שיערוך זה.

באופן פורמלי, אנו מעוניינים לשערך את הפילוג של משך נסיעות המונית בעיר כפונקציית פילוג מצרפי (CDF) או כפונקציית צפיפות הסתברות (PDF).

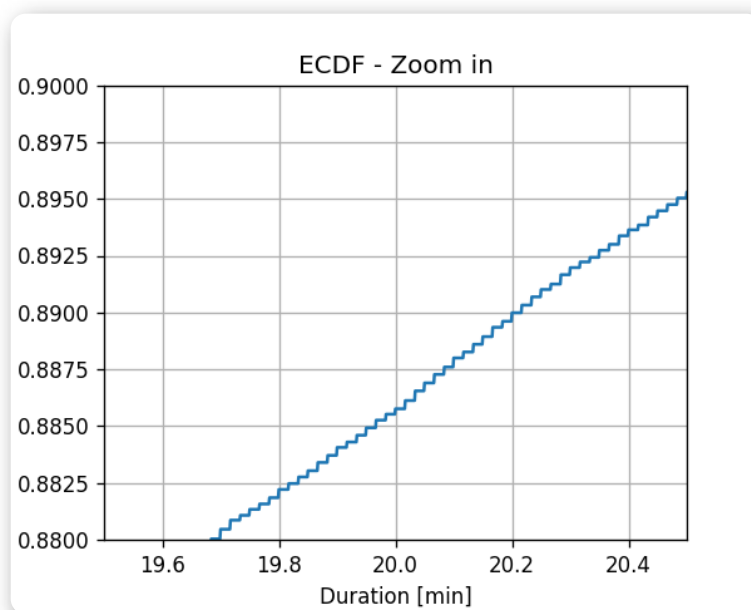
המדגם שלנו לבעיה זו יהיה אוסף משכי הנסיעה מהמדגם הכולל של פרטי הנסיעה. נסמן את המדגם של משך הנסיעה ב  $\{x^{(i)}\}$ .

### שיטה 1: ECDF

נחשב ונשרטט את ה ECDF על פני grid של ערכים בין 0 ל  $\max(\{x^{(i)}\})$  בקפיצות של 0.001:



נסתכל מקרוב על איך נראית פונקציית ה ECDF:



נשים לב שמערך ה ECDF יהיה תמיד מורכב מאוסף של פונקציות מדרגה.

### שאלה

על פי פונקציית הפילוג המצרפי המשוערכת, מהו הסיכוי שנסיעת מונית תערך יותר מ-20 דקות?

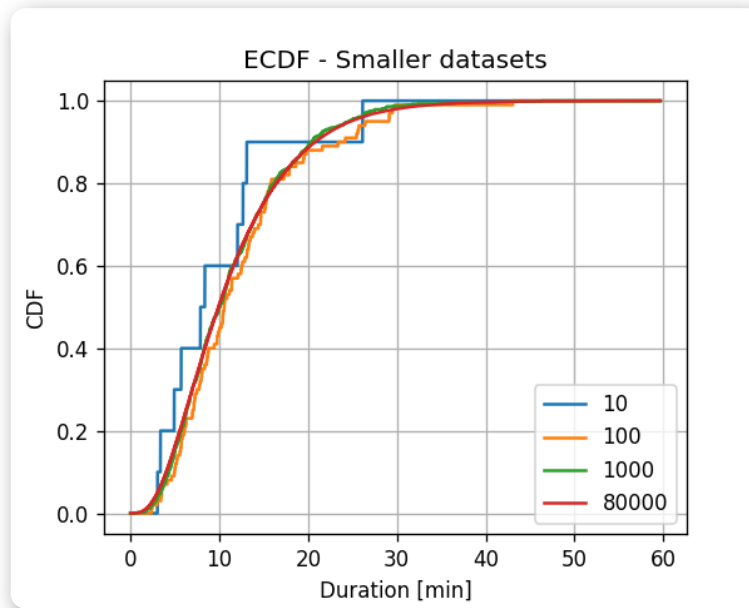
### תשובה

על פי הגדרת הפילוג המצרפי:

$$\Pr(x > 20) = 1 - \Pr(x \leq 20) = 1 - F_x(20) \approx 1 - 0.89 = 0.11$$

### התלות בגודל המדגם

על מנת לראות את התלות של ה ECDF בגודל המדגם נחזור על החישוב עם כמויות קטנות יותר של דגימות במדגם. אנו נבחר בארקאי  $N = 10, 100, 1000, 8000$  דגימות מהמדגם ונחזור על החישוב. התוצאה:



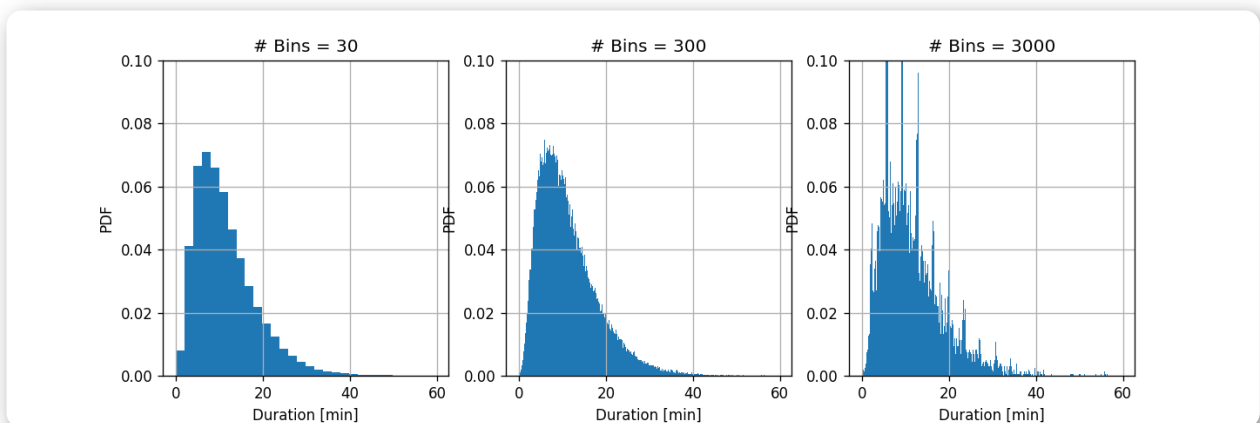
באופן לא מפתיע ניתן לראות כי ככל שאנו מגדילים את מספר הדגימות במדגם המשערך מתקרב יותר ויותר לפונקציה חלקה וניתן גם להראות כי השערך מתקרב (במובן סטטיסטי) לפונקציית הפילוג המצרפי האמיתית.

## שיטה 2: היסטוגרמה

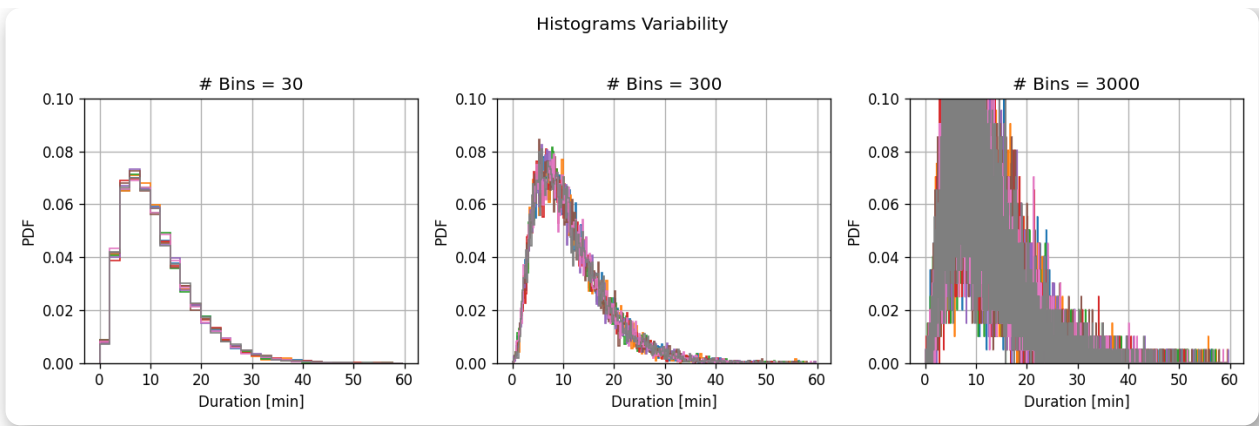
נחשב את ההסטוגרמה של משך הנסיעה בעבור חלוקה של התחום ל-30, 300 ו-3000 תאים.

תזכורת: כלל האצבע לבחירה של מספר התאים הינו  $\sqrt{B} = \sqrt{80000} \approx 280$ .

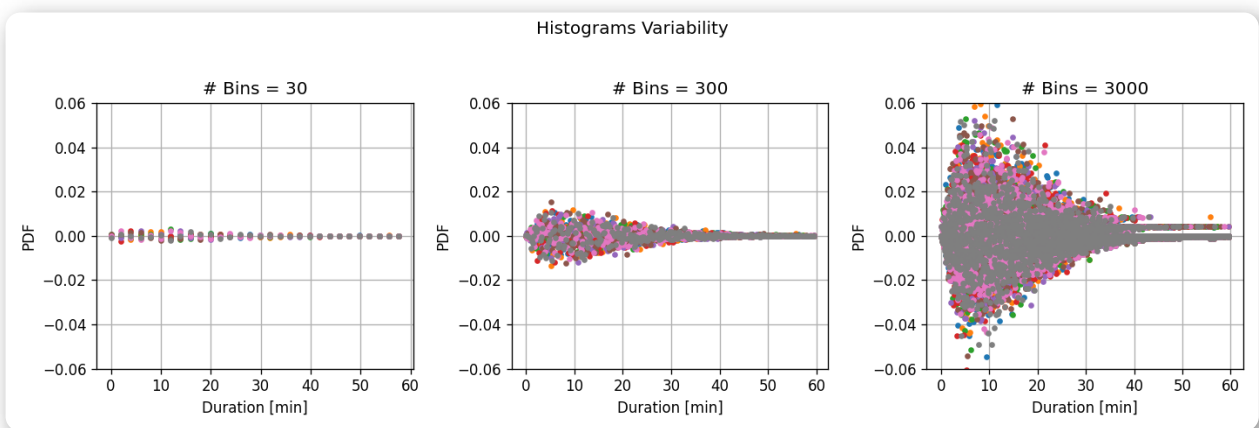
תוצאה:



לפני שנבחן את התוצאות, נריץ מבחן נוסף. ננסה לשערך באופן איכותי את ה variance של כל אחת מההיסטוגרמות. לשם כך נפצל את המגדם ל-8 תתי מדגמים שווים ונחשב היסטוגרמה בעבור כל אחד משמונת תתי המדגם.



בכדי להגדיר את השונות של השיערוך בצורה טובה יותר נחסר משמנות השיערוכים את הממוצע שלהם:



ניתן לראות כי:

- בעבור **מספר גדול של תאים**, ההבדלים בין תתי המדגם השונים (**שונות גדולה**) גדול והתאים צרים ולכן ההיסטוגרמה יכולה לקרב בצורה יותר טובה את פונקציית הצפיפות האמיתית (**הטיה קטנה**)
- בעבור **מספר קטן של תאים**, ההבדלים בין תתי מדגמים שונים קטן (**שונות קטנה**) אך התאים מאד רחבים ולכן לא יכולים לקרב את הפונקציה האמיתית בצורה טובה (**הטיה גדולה**)

זהו למעשה אותו bias-variance tradeoff:

- כאשר **מספר התאים גדול**, כל תא יהיה צר ומקור השגיאה העיקרי ינבע מה**אקראיות** של המדגם הגורמת לשינויים גדולים במספר היחסי של נקודות אשר נופלות בכל תא. שגיאה זו נובעת מה variance של המשערוך. שגיאה זו תלך ותקטן ככל שנגדיל את כמות הדגימות במדגם.
- כאשר **מספר התאים קטן**, מקור השגיאה העיקרי ינבע מ**מיכולת הייצוג המוגבלת** של המודל שלנו. שגיאה זו נובעת מה bias של המשערוך.

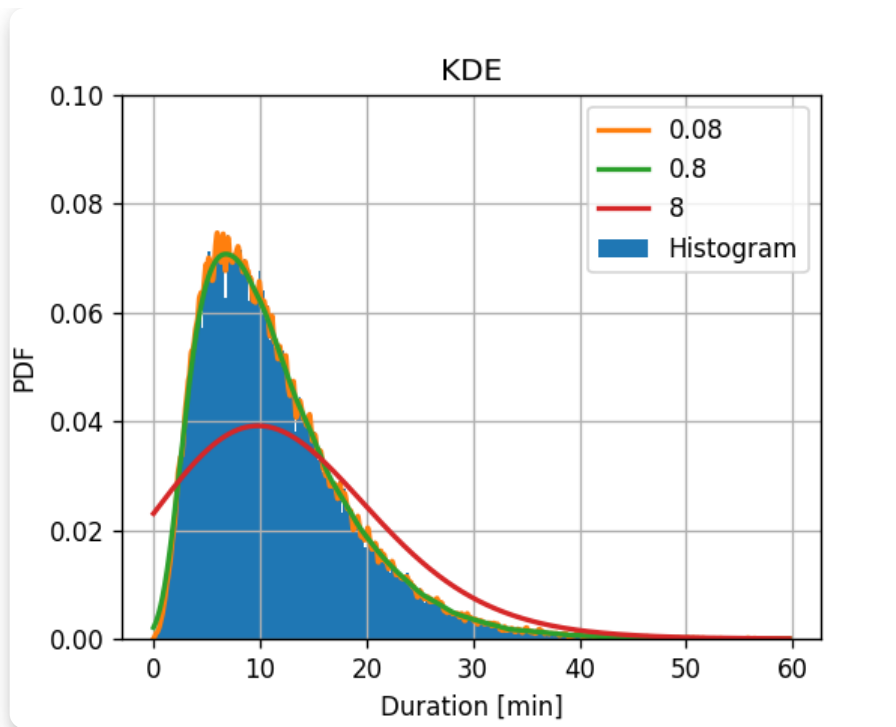
אנו כמובן נשאף לבחור ערך ביניים אשר לא סובל מ variance גדול מידי וגם לא מ bias גדולה מידי. כלל ההאצבע מנסה לתת לעוזר לנו לבחור ערך שכזה.

## שיטה 3: KDE

נשערוך כעת את פונקציית צפיפות ההסתברות בעזרת KDE עם חלון גאוזי. נבחר ערכים שונים לרוחב החלון  $\sigma = 0.08, 0.8, 8$ .

$$\text{זכורת, כלל האצבע מציע לבחור רוחב של: } \sigma = 1.06 \text{ std}(x) N^{-\frac{1}{5}} \approx 0.775$$

לשם השוואה, נשרטט גם את ההיסטוגרמה עם ה 300 תאים:



שוב אנו רואים את ה bias-variance tradeoff:

- עבור בחירה של רוחב צר המשערך יכולה לקרב פרטים "עדינים" יותר, אבל השיערוך רועש יותר. זוהי שגיאת ה variance.
- עבור בחירה של רוחב רחב המשערך מחליק את הפרטים הקטנים, אבל השיערוך פחות רועש יותר. זוהי שגיאת bias.

## בעיית חיזוי: האם נסיעה התרחשה בזמן שעות העבודה

נניח ושעות העבודה ב NYC מוגדרות כשעות שבין 7:00 ו-18:00. נגדיר משתנה אקראי בינארי  $y$  אשר שווה ל 1 אם נסיעה התרחשה בזמן שעות העבודה ו-0 אחרת.

נרצה לבנות חזאי ל  $y$  על סמך  $x$  אשר ימזער את ה **missclassification rate**. נעשה זאת תחת הגישה הגנרטיבית.

נפעל בדומה לתרגיל 5.3. השלבים לפתרון הבעיה:

1. שיערוך הפילוג השולי של  $y$ , זאת אומרת  $\hat{p}_{y,D}(y)$ .
2. שיערוך הפילוג המותנה של  $x$  בהינתן  $y$ , זאת אומרת  $\hat{p}_{x|y,D}(x|y)$ , בעבור כל אחד משני הערכים של  $y$ .
3. בניית החזאי האופטימאלי בהינתן הפילוג המשוערך על פי:  $h(x) = \arg \max_y \hat{p}_{y|x,D}(y|x)$ .

### שלב 1: שיערוך של $\hat{p}_{y,D}(y)$

$y$  הוא משתנה דיסקרטי ולכן השיערוך של הפילוג שלו פשוט:

$$\hat{p}_{y,D}(y) = \frac{1}{N} \sum_{i=1}^N I\{y^{(i)} = y\}$$

נקבל כי:

$$\hat{p}_{y,D}(y) = \begin{cases} 0.51 & y = 1 \\ 0.49 & y = 0 \end{cases}$$

חיזוי עיוור



אם היה ברצוננו לתת חיזוי עיוור (ללא ידיעת  $x$ ) להאם נסיעה התרחשה במהלך שעות העבודה היינו מעוניינים לתת את החיזוי הבא:

$$\hat{y} = \arg \max_y \hat{p}_{y,D}(y) = 1$$

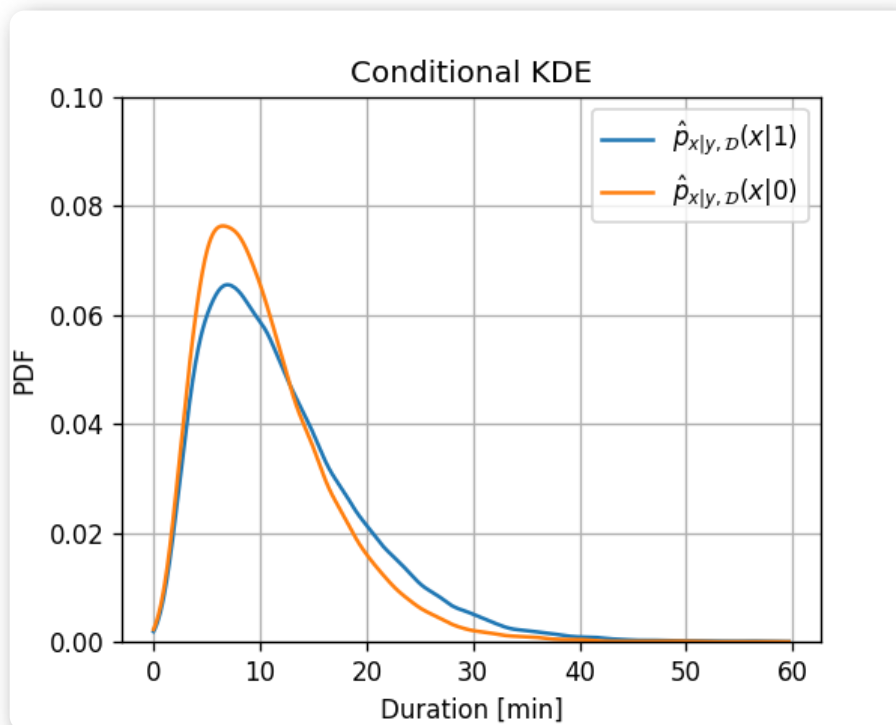
הסיבה שזהו החיזוי האידיאלי נובעת ישירות מן העובדה שיש במדגם יותר נסיעות שהתרחשו בשעות העבודה. שיערכנו שיש סיכוי מעט יותר גדול שנסיעה אקראית תתרחש בשעות העבודה מכיוון שיש לנו סיכוי קטן יותר לטעות בעבור חיזוי זה.

## הערכת ביצועים לחיזוי עיוור

נחשב את ה missclassification rate של החיזוי העיוור (חיזוי קבוע של 1) על ה test set. נקבל את הציון של: 0.49.

## שלב 2: שיערוך $\hat{p}_{x|y,D}(x|y)$

נשתמש פעמיים ב KDE על מנת לשערך את הפילוג המותנה פעם אחת בעבור הדגימות שבהן  $y = 0$  ופעם נוספת בעבור הדגימות שבהן  $y = 1$ :



ניתן לראות כי ישנו שוני קטן בין הפילוגים. לנסיעות מחוץ לשעות העבודה ישנה נטיה קלה יותר לטובת זמני נסיעה קצרים יותר. הבדל קטן זה יעזור לנו לשפר את במעט את יכולת החיזוי שלנו.

## שלב 3: בניית החזאי

עלינו לחשב את:

$$h(x) = \arg \max_y \hat{p}_{y|x,D}(y|x)$$

נתחיל בלהפוך את הפילוג המותנה בביטוי בעזרת חוק בייס על מנת לקבל ביטוי אשר תלוי בפילוגים שחישבנו:

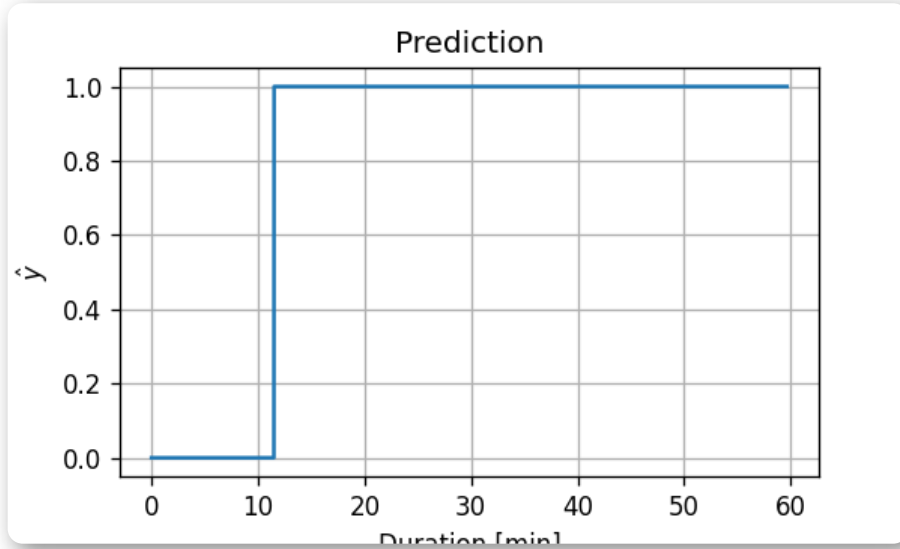
$$= \arg \max_y \frac{\hat{p}_{x|y,D}(x|y)\hat{p}_{y,D}(y)}{\hat{p}_{x,D}(x)}$$

כפי שציינו בתרגיל 5.3, ניתן להפתר מהאיבר במכנה משום שהוא אינו תלוי ב  $y$ :

$$= \arg \max_y \hat{p}_{x|y, \mathcal{D}}(x|y) \hat{p}_{y, \mathcal{D}}(y) = \begin{cases} 1 & \hat{p}_{x|y, \mathcal{D}}(x|1) \hat{p}_{y, \mathcal{D}}(1) > \hat{p}_{x|y, \mathcal{D}}(x|0) \hat{p}_{y, \mathcal{D}}(0) \\ 0 & \text{else} \end{cases}$$

מכאן שהחיזוי יהיה 1 באיזורים שבהם  $\hat{p}_{x|y, \mathcal{D}}(x|1) \hat{p}_{y, \mathcal{D}}(1) > \hat{p}_{x|y, \mathcal{D}}(x|0) \hat{p}_{y, \mathcal{D}}(0)$  ו-0 בכל השאר.

חישוב תנאי זה על פני כל התחום נותן את פונקציית החיזוי הבאה:



מכאן שהחיזוי שלנו יהיה:

$$\hat{y}(x) = \begin{cases} 1 & x \geq 11.4 \\ 0 & \text{otherwise} \end{cases}$$

## הערכת ביצועים

נחשב את ה missclassification rate על ה test set. נקבל את הציון של: 0.46. ציון זה הוא רק מעט יותר טוב מהחיזוי העיוור אשר היה נותן ציון של 0.49. כפי שציינו קודם השיפור הקטן מגיע מההבדלים הקלים שבין שני הפילוגים של הנסיעות בשעות העבודה ומחוצה להן. במקרה זה קיבלנו אומנם שיפור קטן אך ככל שנסתמך בחיזוי שלנו על יותר משתנים השיפורים הקטנים האלו יצברו ונוכל בסוף להגיע לחיזויים מאד מדוייקים.