סיכום הסתברות וסטטיסטיקה

הסתברות

איחוד הסתברויות:

$$P(A) \cup P(B) = P(A) + P(B) - P(A) \cap P(B)$$

חיתוך באירועים בלתי תלויים:

$$P(A) \cap P(B) = P(A) \cdot P(B)$$

משלים:

$$P(\overline{A}) = 1 - P(A)$$

איחוד משלימים:

$$\overline{A} \cap \overline{B} = \overline{A \cup B}$$

$P(B|A)$: B בהנתן ש A

חוק בייס:

$$\begin{aligned} P(A \cap B) =& P(B|A) \cdot P(A) \\ =& P(A|B) \cdot P(B) \end{aligned}$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

נוסחת ההסתברות השלמה:

$$U = \sum_{i=1}^{n} A_i \implies P(B) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

אם נתונים n איברים, אפשר לבחור r איברים מהם $\frac{n!}{(n-r)!}$ פעמים ללא החזרה ועם חשיבות לסדר. או $\binom{n}{r} = \frac{n!}{(n-r)! \cdot r!}$ בלי חשיבות לסדר (מקדם בינומי)

התפלגויות

פונקצית התפלגות: מסכמת את ההסתברות לקבל כל תוצאה אפשרית של משתנה מקרי (משתנה תלוי מקרה).

ניסוי ברנולי: ניסוי עם שתי תוצאות אפשריות, לא בהכרח שוות הסתברות (למשתנה המקרי יש שני ערכים אפשריים).

התפלגות בינומית: התפלגות מספר ההצלחות בניסוי בינומי (סדרה של ניסויי ברנולי). כאשר x מספר ההצלחות בn ניסויים

$$x \sim Bin(n,p)$$
$$P(x=k) = \binom{n}{k} p^k \cdot (1-p)^{n-k}$$

התפלגות הנדסית: התפלגות של מספר הניסויים עד ההצלחה הראשונה (ועד בכלל).

$$x \sim Geo(p)$$
$$P(x=k) = (1-p)^{k-1} p$$

התפלגות פואסון: כאשר θ קצב אירועים, T חלון, x מס' האירועים ב־T

$$x \sim Poiss(\lambda)$$
$$\lambda = \theta T$$
$$P(x=k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

התוחלת μ של משתנה מקרי בדיד x שיכול לקבל את הערכים $x_1, x_2 \cdots$:

$$E(x) = \sum_i x_i \cdot P(x = x_i)$$

תוחלות להתפלגויות השונות:

$$x \sim Bin(n,p) \longrightarrow \mu = np$$
$$x \sim Geo(p) \longrightarrow \mu = \frac{1}{p}$$
$$x \sim Poiss(\lambda) \longrightarrow \mu = \lambda$$

שונות וסטיית תקן: שונות σ^2 היא התוחלת של ההפרשים מהתוחלת וסטיית תקן σ היא שורש השונות

$$V(x) = \sigma^2 = E\left[(x - E(x))^2\right] = E\left(x^2\right) - E^2(x)$$

שונויות להתפלגויות השונות:

$$x \sim Bin(n,p) \longrightarrow \mu = np(1-p)$$
$$x \sim Geo(p) \longrightarrow \mu = \frac{1-p}{p^2}$$
$$x \sim Poiss(\lambda) \longrightarrow \mu = \lambda$$

התפלגות נורמלית

משתנה מקרי סטנדרטי: משתנה מקרי שנובע מהתפלגות עם תוחלת אפס ושונות אחד.
אפשר להפוך משתנה מקרי x למשתנה סטנדאטי y ע"י:

$$Y = \frac{x-\mu}{\sigma} = \frac{1}{\sigma}x - \frac{\mu}{\sigma}$$

פונקציית הצפיפות המצטברת CDF ופונקציית צפיפות ההסתברות PDF: $f(x), F(x)$
ההסתברות שהמשתנה המקרי יהיה נמוך מהערך, והנגזרת של $F(x)$

$$F(x) = P(X < x)$$

$$f(x) = \frac{dF}{dx}$$

התפלגות נורמלית:

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

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

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

ההתפלגות הנורמלית הסטנדרטית:

$$x \sim N(0,1)$$

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

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}}$$

המרה מהתפלגות נורמלית לנורמלית סטנדרטית:

$$Z = \frac{x-\mu}{\sigma}$$

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

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

Z_α: הערך עבורו $P(X < Z_\alpha) = \alpha$

$$\Phi(Z_\alpha) = \alpha$$

אומדנים

אומדן: משתנה מקרי שאומד פרמטר או תכונה של פונקציית ההתפלגות. אומדן הוא לא מוטה אסם התוחלת שלו שווה לפרמטר שהוא אומד $E(\hat{x}=x)$.

ממוצע: אומדן לא מוטה לתוחלת

$$\hat{\mu}=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

משפט הגבול המרכזי: עבור n גדול מספיק, הממוצע של מספר דגימות בלתי תלויות של משתנה מקרי כלשהו בעל תוחלת μ ושונות σ^2 מתפלג נורמלית

$$Y=\frac{1}{n}\sum_{i=1}^{n}x_i$$

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

אומדן לא מוטה לשונות:

$$\hat{V}=s^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2=\frac{n}{n-1}\left(\bar{x^2}-\hat{\mu}^2\right)$$

מדדים לערך מרכזי: ממוצע $\bar{x}$, חציון m (אם n אי־זוגי, $m=x_{\frac{n+1}{2}}$. אם n זוגי, $m=\frac{1}{2}\left(x_{\frac{n}{2}}+x_{\frac{n}{2}+1}\right)$), אמצע הטווח והשכיח. ממוצע ואמצע טווח מושפים מערכים חריגים.

הרחבה לחציון- quantiles: P_α הערך ששבר α מהמשתנים המקריים שיוגרלו מההתפלגות יהיו קטנים ממנו.

$$P_{.25}\equiv Q_1,\qquad P_{.75}\equiv Q_3 \qquad\qquad P_0=-\infty,\qquad P_1=\infty$$

חישוב: בהינתן $(x_1,\,x_2,\,\ldots x_n)$ סדרת ערכים מסודרת
1. נחשב את הquantiles המשויכים לכל ערך: $P_{\frac{i-0.5}{n}}=x_i$
2. ערכי ביניים יחושבו כממוצע משוקלל של הערכים משני הצדדים
3. ערכים קטנים מ$0.5/n$ או גדולים מ$(n-0.5)/n$יקבלו את הערכים x_1(המינימלי) וx_n (המקסימלי) בהתאמה.

מדדים לפיזור: שונות\סטיית תקן σ^2 ו$\mathrm{MAD}=median\left(|x_i-m|\right)$

אומדנים: תוחלת, שונות, ממוצע (אומד לתוחלת) ואומד לשונות הם כולם פעולות ליניאריות, הן מקיימות $f(ax+b)=af(x)+b$ (שונות היא תמיד חיובית)

בדיקת השערות (כללי)

α הסיכוי לדחות את h_0 למרות שהיא נכונה

$$P\left(\mu_0 - M < \bar{x} < \mu_0 + M | h_0\ true\right) = 1 - \alpha$$

עוצמת המבחן β היא הסתברות שלא לדחות את h_0 כאשר היא אינה נכונה.

לחישוב M נשתמש ב$Z_{1-\alpha}$ במבחן חד־צדדי וב$Z_{1-\frac{\alpha}{2}}$ במבחן דו־צדדי.

ערך P (P value): עבור סטטיסטי מסוים, ערך הα הקריטי כך ש $\hat{\theta} = M \pm \theta$. נדחה את השערת האפס h_0 לכל רמת מובהקות גדולה מערך הP. רמת הביטחון גבוהה אם הP value נמוך.

רווח סמך

רווח סמך: כאשר $(1-\alpha)$ הוא רמת הבטחון, θ הוא פרמטר בניסוי וR הוא הרווח (כקבוצה) או M (כרווח לכל כיוון). בין α לM קיים יחס הפוך.

$$P\left(\theta \in R\right) = 1 - \alpha$$

נשתמש בהתפלגות נורמלית כשהשונות ידועה ובהתפלגות t כשהיא לא ידועה.
התפלגות $t\left(\upsilon\right)$: כאשר υ דרגת החופש

$$E\left(t\left(\upsilon\right)\right) = 0$$

$$E\left(t\left(\upsilon\right)\right) = \begin{cases} \infty & 1, 2 \\ \frac{\upsilon}{\upsilon - 2} & \upsilon > 2 \end{cases}$$

רווח סמך לתוחלת:

$$P\left(\hat{\mu} - M < \theta < \hat{\mu} + M\right) = 1 - \alpha$$

בשונות ידועה:

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

רווח סמך לתוחלת אם השונות אינה ידועה ו$n < 30$:

$$M = t_{1-\frac{\alpha}{2}}^{n-1} \cdot \frac{s}{\sqrt{n}}$$

רווח סמך לפרופורציה (ההסתברות להצלחה בניסוי ברנולי):

$$\hat{p} = \frac{k}{n}$$

$$M = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{p\left(1-p\right)}{n}}$$

רווח סמך לשוויון תוחלות:

$$M = Z_{1-\frac{\alpha}{2}} \cdot SD$$

כאשר השונות ידועה:

$$SD = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

כאשר השונות אינה ידועה נשתמש ב $\mathrm{T\ test}$, השיטה הבאה עובדת רק אם אפשר להניח שהשונויות שוות:

$$\hat{\sigma} = \sqrt{\frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}}$$
$$SD = \hat{\sigma}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}} = \hat{\sigma}\sqrt{\frac{n_x + n_y}{n_x n_y}}$$

רווח סמך לשיוויון פרופורציות:

$$\begin{aligned}
\hat{p_x} - \hat{p_y} =& \frac{x}{n_x} - \frac{y}{n_y} \\
\hat{p} =& \frac{n_x \hat{p_x} + y_x \hat{p_y}}{n_x + n_y} = \frac{x + y}{n_x + n_y} \\
M =& \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_x} + \frac{1}{n_y}\right)} \cdot z_{1-\alpha/2}
\end{aligned}$$

שיוויון תוחלות בתצפיות מזווגות:

$$stat = \bar{d} = |x - y|$$

$$M = \frac{s_d}{\sqrt{n}} \cdot z_{1-\alpha/2}$$

התפלגות χ^2

$$X_1 \cdots X_n \sim N(0,1)$$

$$r_n = \sum_{i=1}^{n} X_i^2 \sim \chi^2(n)$$

$$\begin{aligned} E(r_n) &= n \\ V(r_n) &= 2n \end{aligned}$$

מבחן χ^2 לבדיקת התאמה של פונקציית התפלגות

נאסוף $\{O_1 \cdots O_n\}$ ערכים (מספר התצפיות בכל קטגוריה) מהניסוי ונחשב $\{E_1 \cdots E_n\}$ ערכים צפויים לפי השערת האפס (למשל בניסוי של 10 הטלות מטבע $\{E_1 = 5, E_2 = 5\}$). אם באחת הקטגוריות $(E_i < 5) \vee (O_i < 5)$ נאחד את הקטגוריה עם אחרת.
נחשב את הסטטיסטי χ^2 (זה לא ההתפלגות)

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

אם השערת האפס נכונה, אז בקירוב, הסטטיסטי χ^2 מתפלג לפי ההתפלגות χ^2

$$\chi^2 \sim \chi^2(df)$$

כלל הדחייה הוא (זהו מבחן דו צדדי כי החישוב הוא של מרחק):

$$\chi^2 > \chi^2_{1-\alpha}(n-1)$$

מבחן χ^2 לאי תלות (מקרה פרטי של אי התאמה)

H_0 היא שהמשתנים הם בלתי תלויים, H_1 שהם תלויים.
דרגות החופש הם:

$$df = (\#Row - 1) \cdot (\#Column - 1)$$

הסטטיסטי:

$$\chi^2 = \sum_{i,j=1}^{n,m} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

וכלל הדחייה:

$$\chi^2 > \chi^2_{1-\alpha}(n-1)$$