הסקה סטטיסטית- סיכום 2021

הסתברות

הגדרות וחוקים בהסתברות

$P(A|B) = \frac{P(A\cap B)}{P(B)}$ ההסתברות שמאורע A קרה בהינתן שמאורע B קרה. מניחים $P(B) \neq 0$.

- ההסתברות של המאורעות יכולה להתחלף בגדלי הקבוצות, כלומר $P(A|B) = \frac{|A\cap B|}{|B|}$**.**
- בשיטות של עצים: ההסתברויות שכתובות על הצלעות שמובילות לעלים הן כבר ההסתברויות של המאורע העלה בהינתן שכל המאורעות שהובילו עד שם קרו.

Multiplication Rule

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

חוק/ נוסחת ההסתברות השלמה

יהיו A, B_1, B_2, B_3 מאורעות במרחב המדגם Ω, אזי:

$$P(A) = P(A|B_1)\cdot P(B_1) + P(A|B_2)\cdot P(B_2) + P(A|B_3)\cdot P(B_3)$$

מאורעות בלתי תלויים

מאורעות A ו- B ב"ת ⇔ $P(A|B) = P(A)$

$P(B|A) = P(B)$

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

חוק בייס

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

משתנים מקריים בדידים

משתנה מקרי הוא פונקציה $X : \Omega \to R$

- $X = a$ הוא בעצם המאורע $\{w|X(w) = a\}$

probability mass function (pmf)

הפונקציה $p(a) = P(X = a)$

cumulative distribution function (cdf)

הפונקציה $F(a) = P(X \leq a)$

תוחלת של משתנה מקרי בדיד

יהי X משתנה מקרי שיכול לקבל את הערכים $x_1, ..., x_n$. אזי התוחלת של X מסומנת ב- $E(X)$ ומוגדרת כך:

$$E(X) = p(x_1)\cdot x_1 + ... + p(x_n)\cdot x_n = \sum_{i=1}^{n} p(x_i)\cdot x_i$$

<u>תכונות:</u>

1) $E(X+Y) = E(X) + E(Y)$
2) $E(aX+b) = aE(X) + b$

3) $E(h(x)) = \sum_i h(x_i)p(x_i)$

הערה: לפעמים מבקשים לחשב את הממוצע, אז מתכוונים לתוחלת.

שונות

יהי X משתנה מקרי עם תוחלת μ. אזי השונות של X מסומנת ב- $Var(X)$ ומוגדרת כך:

$$Var(X) = E((x-\mu)^2) = \sum_{i=1}^{n} p(x_i)(x_i-\mu)^2$$

חישוב ישיר של השונות: $Var(X) = E(X^2) - (E(X))^2$

<u>תכונות:</u>

1) יהיו a, b קבועים: $Var(aX+b) = a^2 Var(X)$

2) יהיו X, Y משתנים מקריים ב"ת: $Var(X+Y) = Var(X) + Var(Y)$

- כשהמשתנים תלויים צריך להוסיף גם את ה- $Cov(X, Y)$

3) $Var(c) = 0$ לכל קבוע c.

סטיית תקן

יהי X משתנה מקרי עם שונות $Var(X)$. אזי סטיית התקן שלו מסומנת ב- σ ומוגדרת כך: $\sigma = \sqrt{Var(X)}$

<u>התפלגויות מיוחדות:</u>

1. התפלגות ברנולי: משתנה המתפלג ברנולי עם פרמטר p הוא משתנה שיכול לקבל את הערכים 0 או 1 בלבד, ומתקיים: $P(X=0) = 1-p$, $P(X=1) = p$.

   סימון: $X \sim Bernoulli(p)\ or\ Ber(p)$

   - '0' נחשב לכשלון, ו-'1' נחשב להצלחה.
   - פונקציית הסתברות:

   $$\begin{cases} q = 1-p, & k=0 \\ p, & k=1 \end{cases}$$

   - תוחלת: $E(X) = p$
   - שונות: $Var(X) = p(1-p)$

2. התפלגות בינומית: מספר ההצלחות מתוך n ניסויי ברנולי עם פרמטר p, בלתי תלויים.

   סימון: $X \sim Binomial(n,p)\ or\ Bin(n,p)$

   - פונקציית הסתברות: $\binom{n}{k}p^k(1-p)^{n-k}$
   - תוחלת: np
   - שונות: $np(1-p)$

3. התפלגות גיאומטרית: מספר הכשלונות עד שהגענו להצלחה הראשונה בניסויי ברנולי ב"ת.

   סימון: $X \sim geometric(p)\ or\ geo(p)$

   - פונקציית הסתברות: $(1-p)^{k-1}p$
   - תוחלת: $\frac{1}{p}$
   - שונות: $\frac{1-p}{p^2}$

משתנים מקריים רציפים

משתנים המקבלים טווח רציף של ערכים: $(-\infty,\infty),\ [0,\infty),\ [a,b],\ [0,1]$

probability density function (*pdf*)

הפונקציה $P(c \le X \le d) = \int_c^d f(x)\,dx$

- תמיד $f(x) \ge 0$
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$, כלומר: $P(-\infty < X < \infty) = 1$.
- עבור משתנה X שיכול לקבל את הערכים בקטע $[a,b]$ בלבד, ידוע: $\int_a^b f(x)\,dx = 1$

cumulative distribution function (*cdf*)

הפונקציה $F(x) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$

- $0 \le F(x) \le 1$
- הפונקציה לא יורדת
- $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to\infty} F(x) = 1$
- $P(c < X < d) = F(d) - F(c)$
- $F'(x) = f(x)$ (הנגזרת של ה- *cdf* = ה- *pdf*).
- כדי לחשב חציון (*median*) אפשר לבדוק מה ה- q עבורו $F(q) = 0.5$

<u>התפלגויות מיוחדות:</u>

1. התפלגות אקספוננציאלית: בד"כ משתמשים בה כדי למדל זמני המתנה. ערכים בתחום $[0,\infty)$.
   סימון: *exponential*(λ) *or exp*(λ) . כאשר הפרמטר λ אומר "פעם בכמה זמן..."
   - *pdf*: $f(x) = \lambda e^{-\lambda x}$ עבור $0 \le x$
   - דוגמה ל- *likelihood*:

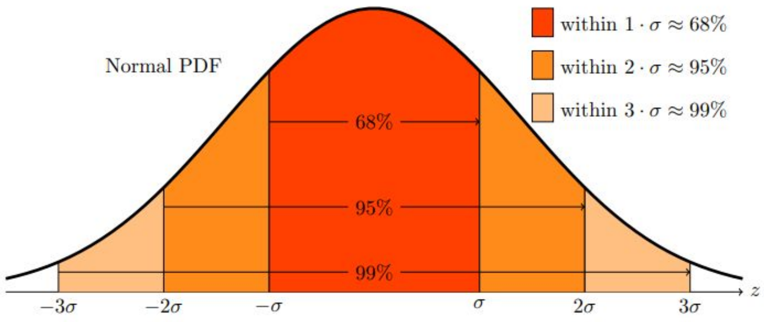
$$f(x_1,\dots,x_5|\lambda) = \prod_{i=1}^{5} \lambda e^{-\lambda x_i} = \lambda^5 e^{-\lambda(x_1+x_2+\cdots+x_5)}$$

2. התפלגות אחידה: טווח הערכים הוא $[a,b]$.
   סימון: $U(a,b)$
   - *pdf*: $f(x) = \frac{1}{b-a}$

3. התפלגות נורמלית: טווח הערכים הוא $(-\infty,\infty)$.
   סימון: $N(\mu,\sigma^2)$
   - *pdf*:

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

- אם מגיעים ל- pdf מהצורה: $f(y) = ce^{-(y-\mu)^2/2\sigma^2}$ אז ידוע ש- $y \sim N(\mu, \sigma^2)$ ואז ידוע שהמקדם $c = \frac{1}{\sigma\sqrt{2\pi}}$ כי הוא מקדם מנרמל.
- משתנה מקרי Z תמיד מתאר משתנה נורמלי סטנדרטי, כלומר $E(Z) = 0,\ Var(Z) = 1$



Rules of thumb:

$P(-1 \le Z \le 1) \approx .68,$
$P(-2 \le Z \le 2) \approx .95,$
$P(-3 \le Z \le 3) \approx .997$

<u>הערה:</u> אם יש X בעל פונקציית pdf מסוימת - $f(x)$, אז ההסתברות ש-X נמצא בין a ל- b מסוימים היא:

$$\int_a^b f(x)\,dx$$

תוחלת

$$E(X) = \int_a^b xf(x)\,dx$$

שונות

$$Var(X) = \int_a^b (x-\mu)^2 f(x)\,dx = E(X^2) - (E(X))^2$$

- השונות של הממוצע הוא כמו ממוצע השונויות. כלומר $Var(\overline{X}) = \frac{Var(X)}{k}$ כאשר k = מספר ה"עותקים" של X, כלומר $X_1, ..., X_k$ הם משתנים בלתי תלויים שהם $identically - distributed\ copies\ of\ X$.

חוק המספרים הגדולים

יהיו $X_1, ..., X_n$ משתנים מקריים בלתי תלויים עם התפלגות זהה (לכולם אותו ממוצע μ ואותה סטיית תקן σ). יהי $\overline{X_n}$ הממוצע שלהם: $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$ (והוא משתנה מקרי בעצמו). אזי לכל מספר קטן יתקיים:

$\lim_{n\to\infty} P(|\overline{X_n} - \mu| < \sigma) = 1$.

המשמעות: אם בוחרים n מספיק גדול, נוכל לקרב מאוד בין הממוצע האמפירי לתוחלת.

היסטוגרמות

סטנדרטיזציה
"נרמול" של משתנה מקרי. יהי משתנה מקרי X עם תוחלת μ ושונות σ^2, אז המשתנה המקרי המייצג את הסטנדרטיזציה של X הוא: $Y = \frac{X-\mu}{\sigma}$.

- ל- Y תמיד יש ממוצע 0 וסטיית תקן 1.
- סטנדרטיזציה של משתנה מקרי נורמלי כלשהו, הופכת אותו למשתנה נורמלי סטנדרטי, שיתואר בד"כ באות Z.

סטנדרטיזציה לממוצע $\bar{x}$: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ כאשר כל ה- x_iים נדגמו מהתפלגות $N(\mu,\ \sigma^2)$

משפט הגבול המרכזי
יהיו $X_1, ..., X_n$ משתנים מקריים בלתי תלויים עם ממוצע μ וסטיית תקן σ. לכל n נגדיר את הסכום - $S_n = X_1 + X_2 + ... + X_n$, ואת הממוצע האמפירי - $\overline{X_n} = \frac{1}{n}(X_1 + X_2 + ... + X_n)$.
אזי עבור n מספיק גדול: $\overline{X_n} \approx Normal(\mu,\ \frac{\sigma^2}{n})$

$$S_n \approx Normal(n\mu,\ n\sigma^2)$$

<u>הערה:</u> כשכתוב $i.i.d$ הכוונה היא למשתנים מקריים בלתי תלויים עם התפלגות זהה (אותה תוחלת ושונות).

התפלגויות משותפות

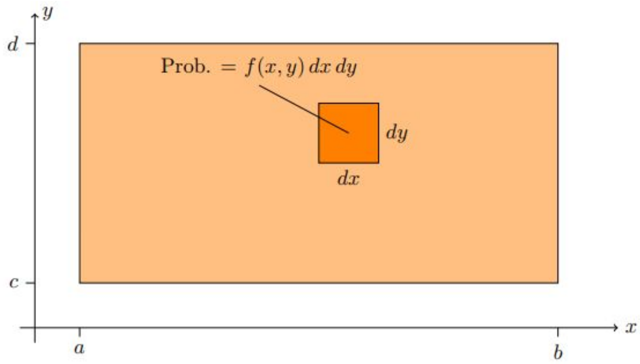
המקרה הבדיד:
יהיו X, Y משתנים מקריים בדידים, כאשר X מקבל את הערכים $\{x_1, ..., x_n\}$, ו- Y מקבל את הערכים $\{y_1, ..., y_m\}$. הזוג הסדור (X, Y) מקבל ערכים מתוך $\{(x_1, y_1),\ (x_1,\ y_2),\ ...,\ (x_n,\ y_m)\}$.
ההתפלגות המשותפת (joint pmf) של X ו- Y היא הפונקציה $p(x_i, y_j)$, שנותנת את ההסתברות לתוצאה המשותפת: $X = x_i,\ Y = y_j$

joint probability table

$X\backslash Y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m
x_1	$p(x_1, y_1)$	$p(x_1, y_2)$	$\cdots$	$p(x_1, y_j)$	$\cdots$	$p(x_1, y_m)$
x_2	$p(x_2, y_1)$	$p(x_2, y_2)$	$\cdots$	$p(x_2, y_j)$	$\cdots$	$p(x_2, y_m)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_i	$p(x_i, y_1)$	$p(x_i, y_2)$	$\cdots$	$p(x_i, y_j)$	$\cdots$	$p(x_i, y_m)$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	
x_n	$p(x_n, y_1)$	$p(x_n, y_2)$	$\cdots$	$p(x_n, y_j)$	$\cdots$	$p(x_n, y_m)$

ה- *joint pmf* חייבת לקיים שני תנאים:
(1 $0 \le p(x_i,\ y_j) \le 1$
(2 סכום ההסתברויות = 1, כלומר: $\sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,\ y_j) = 1$

המקרה הרציף:
אותו דבר, רק שמחליפים את הקבוצה הדיסקרטית של הערכים באינטרוול רציף, את ה- *pmf* ב- *pdf*, ואת הסכום באינטגרלים.
יהיו X, Y משתנים מקריים בדידים, כאשר X מקבל ערכים בתחום $[a, b]$, ו- Y מקבל ערכים בתחום $[c, d]$, אז (X, Y) מקבל ערכים בתחום $[a, b] \times [c, d]$.
ה- *pdf* המשותף של X ו- Y היא הפונקציה $f(x, y)$ שנותנת את ה- *probability density* ב- (x, y). כלומר $f(x, y)\,dx\,dy$ מייצג את ההסתברות ש- (X, Y) נמצא במלבן קטן עם רוחב dx וגובה dy סביב (x, y):



ה- $joint\ pdf$ חייבת לקיים שני תנאים:

3) $0 \le f(x,y)$

4) "סכום" ההסתברויות = 1, כלומר: $\int_c^d \int_a^b f(x,y)\ dx\ dy = 1$

<u>הערה:</u> ניתן לקבל כאן ערכים גדולים מ-1, כי זו פונקציית צפיפות, ולא פונקציית הסתברות!

גם במקרה הבדיד, וגם במקרה הרציף, יש להם cdf משותף: $F(X,Y) = P(X \le x,\ Y \le y)$

במקרה הבדיד: זה שווה ל- $\sum_{x_i \le x} \sum_{y_j \le y} p(x_i, y_j)$

במקרה הרציף: זה שווה ל- $\int_c^y \int_a^x f(u,v)\ du\ dv$

<u>תכונות:</u>

1) $F(X,Y)$ היא פונקציה לא יורדת

2) באזור השמאלי-הנמוך של הטווח: $F(x,y) = 0$

3) באזור הימני-הגבוה של הטווח: $F(x,y) = 1$

משתנים מקריים בלתי תלויים

X ו- Y משתנים מקריים ב"ת אם $F(X,Y) = F_X(X)F_Y(Y)$

- בדידים: X ו- Y ב"ת אם $p(x_i, y_i) = p_X(x_i)p_Y(y_j)$
- רציפים: X ו- Y ב"ת אם $f(x,y) = f_X(x)f_Y(y)$

שונות משותפת - $Covariance$

- מודד את ההתפלגות המשותפת של X, Y
- מודד קשר לינארי בין המשתנים $Cov(X,Y) = E((X - E(X))(Y - E(Y))$
- אם X ו- Y ב"ת אז $Cov(X,Y) = 0$, אבל הגרירה היא חד כיוונית!

<u>תכונות השונות המשותפת:</u>

- $Cov(aX + b,\ cY + d) = acCov(X,Y)$
- $Cov(X_1 + X_2,\ Y) = Cov(X_1, Y) + Cov(X_2, Y)$
- $Cov(X,X) = Var(X)$
- $Cov(X,Y) = E(XY) - E(X)E(Y)$

קורולציה (עד כמה המשתנים מתואמים)

- דומה ל- $Covariance$ רק שמורידים את ה-scale:

$Cor(X,Y) = \rho\frac{Cov(X,Y)}{\sigma_X\sigma_Y}$ כאשר תמיד $-1 \le \rho \le 1$

<u>סטטיסטיקה</u>
סטטיסטיקה בייסיאנית

הגדרות:

$Statistic$: כל מה שיכול להיות מחושב ע"פ הדטה שלנו (בסטטיסטיקה מקבלים דטה וצריכים להסיק לפיו מסקנות). חייב להיות $Observable$.
ה- $statistic$ לא יכול להיות תלוי בערך האמיתי של פרמטר לא ידוע, אבל כן יכול להיות תלוי בערך משוער שלו הוא משתנה מקרי, כי כל פעם שנעשה ניסוי חדש נקבל דטה חדש, שיניב תוצאות ומסקנות חדשות.
$Point\ Statistic$: ערך יחיד שאנו מחשבים על הדטה (לדוגמה ממוצע).
$Interval\ Statistic$: קטע $[a,b]$ המחושב מהדטה.

$Likelihood$
נסמן: $H=$ $hypothesis$, ההיפותזה שלנו
$D=$ $data$, הדטה שלנו
חוק בייס נותן את ההסתברות של ההיפותזה בהינתן הדטה: $P(H|D)=\frac{P(D|H)\cdot P(H)}{P(D)}$
$P(H|D)$ נקרא "$posterior$"
$P(H)$ נקרא "$prior$"
$P(D|H)$ נקרא "$likelihood$"

- את $P(D)$ נחשב בד"כ על פי נוסחת ההסתברות השלמה. נקרא "total probability of the data" או "the prior predictive probability of the data" .

MLE ($Maximum\ Likelihood\ Estimate$)
דרך לאמוד את הערך של פרמטר בו אנו מתעניינים.
הערך שממקסם את ה- $likelihood$. כלומר הערך שאם נציב אותו ב- p (ההיפותזה), נקבל את ה- $likelihood$ המקסימלי. עושים זאת ע"י מציאת נקודת מקסימום באופן הרגיל - גזירת ה- $likelihood$ והשוואה ל-0.

- הרבה פעמים יותר נוח להשתמש ב- $log\ likelihood$ על מנת למצוא את ה- MLE (לא ישתנה).
בד"כ משתמשים ב- ln , שכאשר מפעילים אותו על הנגזרת שהיא מכפלת איברים, הכפל ביניהם הופך לחיבור, ומוסיפים לכל אחד מהם ln .

$Bayesian\ Updating$
התהליך של מעבר מה- $prior$ ל- $posterior$.
משתמשים ב- $data$ שנוסף לנו על מנת לעדכן את ההסתברויות לכל היפותזה.
$bayesian\ update\ table$

hypoth.	prior	likeli.	Bayes numerator	posterior
θ	$f(\theta)\,d\theta$	$f(x\mid\theta)$	$f(x\mid\theta)f(\theta)\,d\theta$	$f(\theta\mid x)\,d\theta=\frac{f(x\mid\theta)f(\theta)\,d\theta}{f(x)}$
total	1		$f(x)$	1

$$f(x)=\int f(x\mid\theta)f(\theta)\,d\theta$$

$Probabilistic\ Prediction$ (או $Probabilistic\ forecasting$)
הרעיון הוא להקצות הסתברות לכל תוצאה אפשרית של ניסוי עתידי.

Odds

ה-$odds$ של מאורע הם: $O(E)=\frac{P(E)}{P(E^C)}$

- בד"כ עבור 2 בחירות: E ו- $not\ E$.
- ניתן לחלק כך כמה תוצאות אפשריות לשתי קבוצות.
- עבור מאורע E עם הסתברות p: $O(E)=\frac{p}{1-p}$
- אם מאורע הוא "נדיר" ($rare$), אז $P(E)\approx O(E)$. ההסתברות שלו נמוכה.

Posterior Odds

$$O(M|F)=\frac{P(M|F)}{P(M^C|F)}=\frac{P(F|M)}{P(F|M^C)}\cdot\frac{P(M)}{P(M^C)}$$

כאשר: $Posterior\ Odds\ =\ O(M|F)$

$Bayes\ Factor\ =\ \frac{P(F|M)}{P(F|M^C)}$

$Prior\ Odds\ =\ \frac{P(M)}{P(M^C)}$

מרחב השערות רציף

$prior$ = ההסתברות של משהו לקרות לפני שרואים $data$.

$posterior$= ההסתברות לאחר שרואים $data$.

נוסחת ההסתברות השלמה

- עבור משתנים בדידים:

יהיו $H_1, H_2, ..., H_n$ השערות, ויהי D ה-data, אזי: $P(D)=\sum_{i=1}^{n}P(D|H_i)P(H_i)$

- עבור משתנים רציפים:

תהי השערה θ בטווח $[a,b]$ ומשתנה מידע בדיד X, אזי: $P(X)=\int_a^b P(x|\theta)f(\theta)\,d\theta$

- נקרא גם $Prior\ Predictive\ Probability$ של תוצאה x.

חוק בייס למרחב רציף

יהי θ משתנה רציף עם pdf $f(\theta)$, וטווח $[a,b]$, ויהי X משתנה מקרי המייצג data.

ה- $likelihood$: $P(X|\theta)$

חוק בייס: $f(\theta|X)\,d\theta=\frac{P(X|\theta)f(\theta)d\theta}{P(X)}=\frac{P(X|\theta)f(\theta)d\theta}{\int_a^b P(X|\theta)f(\theta)d\theta}$

$f(\theta|X)\,d\theta$ נקרא "$posterior\ pdf$".

- כשעושים מספר ניסויים, למשל הטלות, אז ה $prior$ של ההטלה השנייה הוא ה-$posterior$ של ההטלה הקודמת. זה נקרא $posterior\ predictive\ probability$.

דוגמה לפתרון שאלה:

(b) We are asked for posterior predictive probabilities. Let x be the value of the next roll. We have to compute the total probability

$$p(x|\text{data}) = \sum p(x|H)p(H|\text{data}) = \sum \text{likelihood} \times \text{posterior}.$$

The sum is over all hypotheses. We can organize the calculation in a table where we multiply the posterior column by the appropriate likelihood column. The total posterior predictive probability is the sum of the product column.

hyp.	posterior to data	likelihood (i) $x = 5$	post. to (i)	likelihood (ii) $x = 15$	post. to (ii)
H_4	0	0	0	0	0
H_6	0.243457	1/6	0.04058	0	0
H_8	0.684723	1/8	0.08559	0	0
H_{12}	0.060864	1/12	0.00507	0	0
H_{20}	0.010956	1/20	0.00055	1/20	0.00055
Tot.	0.22819		0.13179		0.00055

So, (i) $p(x = 5|\text{data}) = 0.132$ and (ii) $p(x = 15|\text{data}) = 0.00055$.

- ה- *posterior pdf* בשני המקרים הבאים שקול:
  1) נתונים **מספר** ההצלחות ומספר הכשלונות מתוך מספר ניסויים
  2) נתון הסדר של התוצאות שהתקבלו

  ההבדל היחיד ביניהם הוא שהמקדם הבינומי של ה- *likelihood* במקרה הראשון הוא בחירת k הצלחות מתוך n ניסויים, ואילו במקרה השני הוא פשוט 1 (כי לא צריך לבחור, זה ידוע).

- כשמבקשים את ה- *posterior distribution* (על θ למשל) רוצים בעצם את ה- $f(\theta|D)$, הצפיפות:

$$f(\theta\,|, \text{data}) = \frac{19!}{10!\,8!}\theta^{10}(1-\theta)^8$$

וכשמבקשים את האינטגרל של ה- *posterior predictive probability* להצלחה בניסוי הבא:

The law of total probability says that the posterior predictive probability of success is

$$P(\text{success}\,|\,\text{data}) = \int_0^1 f(\text{success}\,|\,\theta)\cdot f(\theta\,|\,\text{data})\,d\theta$$
$$= \int_0^1 \theta\cdot\frac{19!}{10!\,8!}\theta^{10}(1-\theta)^8\,d\theta = \int_0^1 \frac{19!}{10!\,8!}\theta^{11}(1-\theta)^8\,d\theta$$

התפלגות *Beta*

התפלגות *beta*: $beta(a, b)$ היא התפלגות בעלת שני פרמטרים עם טווח $[0, 1]$ ו- *pdf*:

$$f(\theta) = \frac{(a+b-1)!}{(a-1)!(b-1)!}\theta^{a-1}(1-\theta)^{b-1}$$

- אם מגיעים להתפלגות מהצורה $c\cdot\theta^{a-1}(1-\theta)^{b-1}$ אז יודעים בוודאות שהמקדם הוא המקדם המתאים להתפלגות *beta*, כי הוא היחיד שמנרמל את זה.

Conjugate priors

כעת ה- *data* הוא רציף, וה- *likelihood* הוא *pdf*.

נאמר ש- *prior* הוא *conjugate* ל- *likelihood* מסוים אם לאחר שכופלים אותו ב- *likelihood* מקבלים *posterior* מאותה צורה של ה- *prior*.

נוסחאות:

עבור עדכון של *Normal prior* ו- *Normal likelihood*:

$$a = \frac{1}{\sigma^2_{prior}},\ b = \frac{n}{\sigma^2},\ \mu_{post} = \frac{a\mu_{prior}+b\bar{x}}{a+b},\ \sigma^2_{post} = \frac{1}{a+b}$$

$\sigma_{prior},\ \mu_{prior}$ אמורים להיות נתונים בד"כ, כי נתונה ההתפלגות של המשתנה (למשל θ).

σ = ההתפלגות ממנה דגמנו.

n = מספר הדגימות.

$\bar{x}$ = ממוצע הדגימות.

$\sigma_{post},\ \mu_{post}$ הם הנתונים של התפלגות ה- *posterior*. לפעמים כשמבקשים למצוא את ה- *posterior pdf* צריך בעצם למצוא אותם (ומה ההתפלגות שלו, למשל נורמלית עם הפרמטרים האלה).

לאחר שמחשבים את ה- *posterior pdf*: $f(\theta|data) \sim N(\mu_{post}, \sigma^2_{post})$, ניתן למצוא את ה- *probability interval* בעזרת הנוסחה: $[\mu_{post} - z_{\alpha/2}\sigma,\ \mu_{post} + z_{\alpha/2}\sigma]$

הערות:

- *Normal likelihood* זה ההתפלגות של ה-data. הפרמטר σ מתייחס אליו.
- *Normal prior* ו- *Normal Posterior* מתכוונים להתפלגות של הפרמטר המבוקש (למשל θ). הראשון - מה שמניחים בהתחלה, לפני הוספת ה-data, והשני - העדכון לאחר הוספת ה-data.

עבור התפלגות *Beta*:

התפלגות *beta* היא *conjugate prior* עבור התפלגות בינומית (וברנולי). זה אומר שאם פונקציית ה- *likelihood* היא בינומית, וה- *prior distribution* הוא *beta*, אז ה- *posterior* הוא גם *beta*.

hypothesis	data	prior	likelihood	posterior
θ	x	$\text{beta}(a, b)$	$\text{binomial}(N, \theta)$	$\text{beta}(a + x, b + N - x)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$c_2\theta^x(1-\theta)^{N-x}$	$c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

hypothesis	data	prior	likelihood	posterior
θ	x	$\text{beta}(a, b)$	$\text{Bernoulli}(\theta)$	$\text{beta}(a + 1, b)$ or $\text{beta}(a, b + 1)$
θ	$x = 1$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	θ	$c_3\theta^{a}(1-\theta)^{b-1}$
θ	$x = 0$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$1-\theta$	$c_3\theta^{a-1}(1-\theta)^{b}$

כאשר הקבועים הם:

$$c_1 = \frac{(a+b-1)!}{(a-1)!\,(b-1)!} \qquad c_2 = \binom{N}{x} = \frac{N!}{x!\,(N-x)!} \qquad c_3 = \frac{(a+b+N-1)!}{(a+x-1)!\,(b+N-x-1)!}$$

התפלגות *beta* היא גם *conjugate prior* עבור התפלגות גיאומטרית.

hypothesis	data	prior	likelihood	posterior
θ	x	$\text{beta}(a, b)$	$\text{geometric}(\theta)$	$\text{beta}(a + x, b + 1)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$\theta^x(1-\theta)$	$c_3\theta^{a+x-1}(1-\theta)^{b}$

טבלה מסכמת:

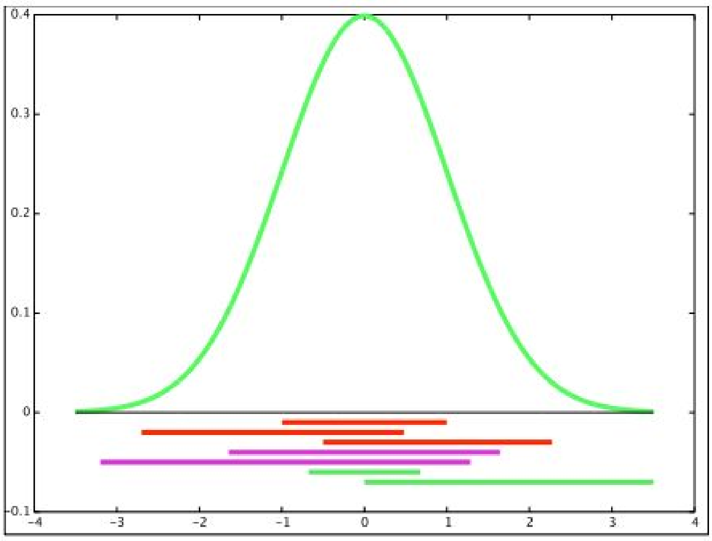
	hypothesis	data	prior	likelihood	posterior
Bernoulli/Beta	$\theta \in [0,1]$	x	$\text{beta}(a,b)$	$\text{Bernoulli}(\theta)$	$\text{beta}(a+1,b)$ or $\text{beta}(a,b+1)$
	θ	$x=1$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	θ	$c_3\theta^{a}(1-\theta)^{b-1}$
	θ	$x=0$	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$1-\theta$	$c_3\theta^{a-1}(1-\theta)^{b}$
Binomial/Beta	$\theta \in [0,1]$	x	$\text{beta}(a,b)$	$\text{binomial}(N,\theta)$	$\text{beta}(a+x,b+N-x)$
(fixed N)	θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$c_2\theta^{x}(1-\theta)^{N-x}$	$c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$
Geometric/Beta	$\theta \in [0,1]$	x	$\text{beta}(a,b)$	$\text{geometric}(\theta)$	$\text{beta}(a+x,b+1)$
	θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$\theta^{x}(1-\theta)$	$c_3\theta^{a+x-1}(1-\theta)^{b}$
Normal/Normal	$\theta \in (-\infty,\infty)$	x	$\text{N}(\mu_{\text{prior}},\sigma^2_{\text{prior}})$	$\text{N}(\theta,\sigma^2)$	$\text{N}(\mu_{\text{post}},\sigma^2_{\text{post}})$
(fixed σ^2)	θ	x	$c_1\exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$	$c_2\exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$	$c_3\exp\left(\frac{(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$

- בהתפלגות נורמלית, ככל שנקבל עוד דטה, השונות תקטן תמיד.
- בהתפלגות $beta-binomial$ אם נקבל עוד דטה, השונות בדרך כלל תקטן, אבל היא עשויה גם לגדול.

Probability Interval

אם $P(a \le \theta \le b) = p$ אזי האינטרוול $[a,b]$ מייצג $p - Probability\ Interval$ עבור θ.
במילים אחרות: האינטרוול אומר מה ההסתברות ש-θ נמצאת בין a ל-b.

- דוגמה: בין כל $quantile\ x$ לבין $quantile\ y = x + 0.5$ יש $0.5\ probability\ interval$. נקרא גם $50\%\ probability\ interval$.
- סימונים ($Q - notation$): נסמן ב-q_p את ה-$quantile$ ה-p, ואז: $[q_{0.25}, q_{0.75}],\ [q_{0.05}, q_{0.55}]$ הם $0.5\ probability\ intervals$.
- ה-$p - probability\ interval$ של ה-$posterior$ בד"כ קטן יותר מזה של ה-$prior$, הודות ל-data.
- $Symmetric\ probability\ interval$ הוא אינטרוול סימטרי סביב ה-0.
- עבור התפלגות נורמלית, יש $Probability\ Intervals$ ידועים:



red = 0.68, magenta = 0.9, green = 0.5

frequentist statistics

כאן נעבוד רק עם ה-$likelihood$, כי אין "הסתברות לפרמטר", ולכן אין שימוש ב-$prior$.

! ההבדל הגדול בין השיטה הזאת לבין $Bayesian\ Statistic$:

- כאשר ה-$prior$ ($P(H)$) ידוע, כולם משתמשים בחוק בייס ומחשבים את ה-$posterior$.

ההבדל מגיע כאשר ה-$prior$ לא ידוע:

- ע"פ השיטה הבייסיאנית, חייב $prior$, לכן צריך לפתח $subjective\ prior$ כלשהו מהמידע הטוב ביותר שיש לנו.
- לעומת זאת, ע"פ השיטה ה-frequentist-ית, נוכל להסיק מסקנות על סמך ה-$likelihood$ בלבד.

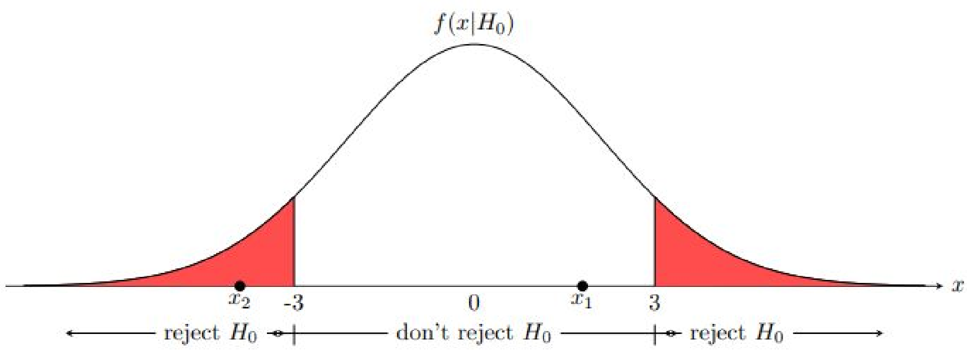
<u>הערה:</u> ה-$Frequentist\ methods$ נותנות הסתברויות ל-data תחת השערות, מה שאומר שלא ניתן לדעת את ההסתברויות. ה-odds של ההשערות הללו!

מבחני השערות

H_0 = השערת האפס. מצב קיים.
H_1/H_A = השערה אלטרנטיבית. השערת החוקר (מה שהוא יצטרך להוכיח, בעזרת מדגם).
$test\ statistic$ = x
$rejection\ region$ = האזור שאם ה-$statistic$ המחושב נופל בו, ניתן לדחות את השערת האפס.
$p(x|H0)$ או $f(x|H0)$ נקראים "$null\ distribution$".

הרעיון הוא לגלות האם אפשר לדחות את השערת האפס לטובת ההשערה האלטרנטיבית.
שלבים:

- איסוף $data$ שרלוונטי להיפותזה שלנו, בעזרת מדגם מהאוכלוסיה.
- חישוב $statistic$ (למשל ממוצע, $\overline{x}$) על הדטה הזה, עם $null\ distribution$ ידועה ($f(\overline{x}|H_0)$).
- מציירים עקומת פעמון, כאשר באמצע שלה שמים את הערך של השערת האפס.
- אם ההשערה האלטרנטיבית היא חד צדדית (גדול מהשערת האפס או קטן מהשערת האפס), אז מקציבים שטח קטן בקצה הימני או השמאלי של העמודה (בהתאמה), ואם ההשערה היא דו צדדית ($\neq$ מהשערת האפס, כלומר לשני הכיוונים), אז מקציבים שטח בשני הקצוות.
- השטח הזה נקרא "רמת מובהקות", או $rejection\ region$, ומסומן ב-α.
- בודקים האם ניתן לדחות את השערת האפס, כך: מגדירים השערה אלטרנטיבית H_A, אם ה-$statistic$ שחישבנו על המדגם נופל בשטח המסומן (אזור רמת המובהקות α), אז ניתן לדחות את השערת האפס - החוקר צודק. אחרת, לא ניתן לדחות את השערת האפס, החוקר טועה.
- מחשבים את ה-$power$ בעזרת $f(\overline{x}|H_A)$



נקודה קריטית (*critical value*) - הנקודה שמפרידה בין רמת המובהקות לשאר השטח מתחת לעקומה. מוצאים אותה בעזרת הטבלאות של Z, T, χ^2 שנתונות במבחן.

Significant Level

ההסתברות ש- x ייפול באיזור דחייה, בהינתן H_0. אם זה בדיד זה סכום ההסתברויות של הערכים באזור הדחייה. מסומן ב- α.

Z Test

ההנחה היא שה- *data* מתפלג נורמלית. μ לא ידוע, σ ידוע.
כדי לבדוק האם ה *statistic* של המדגם נמצא ב- *rejection region* או לא, צריך למצוא את הערך שמתאים לו בטבלת Z, בעזרת הנוסחה: $z = Z_{\bar{x}} = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$, כאשר μ_0 = הערך של השערת האפס. הממוצע/ תוחלת.
הערך של p (*p value*) יהיה $p = P(Z > z \mid H_0)$ במקרה של *right side p − value*
$p = P(|Z| > z \mid H_0)$ במקרה של *two sided p − value* .
significant level : אם $p \leq \alpha$ נדחה את H_0 לטובת H_A, ואם $p > \alpha$ לא נדחה.

! ההבדל בין ה- *Significant Level* ל- *p value* :
הראשון מתייחס להסתברויות באזור הדחייה, והשני מתייחס להסתברויות באזורים שאחרי ה- z שחישבנו (מהנתונים). כלומר $p-value$ הוא ההסתברות לקבל תוצאה קיצונית לפחות כמו z שקיבלנו.

Z Table:

- נותנת לכל ערך את ה- *CDF* , כלומר ההסתברות שהערך קטן מהערך הקריטי. מ- $-\infty$ עד הערך.
- בגלל שזו הסתברות, כל השטח מ- $-\infty$ עד ∞ = 1.
- ה- *null distribution* היא $N(0, 1)$, בגלל זה הכל סימטרי!

מינוחי השערות

השערה פשוטה = השערה שניתנת באופן ישיר, ידוע מה ה- θ למשל. התפלגות הדגימה מוגדרת במלואה.
השערה מורכבת = השערה שלא ניתנת באופן ישיר, למשל אומרים ש- $\theta > 0.4$ (ישנם כמה ערכים אפשריים לפרמטר בו אנו מתעניינים).

סוגי טעויות

		True state of nature	
		H_0	H_A
Our decision	Reject H_0	Type I error	correct decision
	'Don't reject' H_0	correct decision	Type II error

Type I: false rejection of H_0
Type II: false non-rejection ('acceptance') of H_0

Significant Level = Type I Error = False Positive = ההסתברות שה- *test statistic* ב- *rejection region* בהינתן ש- H_0 נכונה (דחינו את H_0 בטעות).
Power = 1 − *Type II Error* = *True Positive* = ההסתברות שה- *test statistic* ב- *rejection region* בהינתן ש- H_A נכונה (דחינו את H_0 כשהיה צריך).

- H_A קובעת את ה- *Power* של הניסוי.
- *Significant Level* ו- *Power* שניהם הסתברויות של ה- *rejection region* .

- המטרה שלנו היא למזער את ה $Significant\ Level$ (שיהיה קרוב ל-0), ולמקסם את ה $Power$ (קרוב ל-1).

$T\ Test$

כאן מניחים שה- $data$ מתפלג נורמלית אך גם μ וגם σ לא ידועים. כלומר: $x_1, ..., x_n \sim N(\mu, \sigma^2)$

$T\ Table$:

- נותנת לכל ערך את ההסתברות שהערך גדול מהערך הקריטי. כלומר מהערך הקריטי עד ∞.
- גם כאן כל השטח שווה ל-1
- גם כאן הכל סימטרי לשני הכיוונים!

$One\ Sample\ T-Test$

הנוסחה של $t\ test\ statistic$: $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, כאשר μ_0 הוא ה- $Null\ hypothesis$ עבור הערך של μ [מניחים $\mu = \mu_0$].

$Sample\ variance$:

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ כאשר n = גודל המדגם, $\bar{x}$ = ממוצע התוצאות, x_iים = התוצאות עצמן.

- כדי למצוא את סטיית התקן, מוציאים לזה שורש.
- ככל שה- n גדל, ה- $t-test$ הולך ושואף להתפלגות נורמלית.

$Null\ distribution$: $f(t|H_0)$ הוא ה- pdf של $T \sim t(n-1)$, כלומר התפלגות t עם $n-1$ דרגות חופש.

$Two\ sided\ p-value$: $p = P(|T| > |t|)$

$Two\ Sample\ T-Test$

מניחים שיש 2 סטים של $data$ המתפלגים נורמלית עם σ זהה: $x_1, ..., x_n \sim N(\mu_x, \sigma^2)$

$y_1, ..., y_m \sim N(\mu_y, \sigma^2)$

ה- $Null\ hypothesis$ הוא $\mu_x = \mu_y$

נוסחת ה- $pooled\ variance$: $s_p{}^2 = \frac{(n-1)s_x{}^2+(m-1)s_y{}^2}{n+m-2}(\frac{1}{n} + \frac{1}{m})$

$t\ test\ statistic$: $t = \frac{\bar{x}-\bar{y}}{s_p}$

$Null\ distribution$: $f(t|H_0)$ הוא ה- pdf של $T \sim t(n+m-2)$

$\chi^2\ Test$

ה- $test\ statistic$: $\chi^2\ statistic$

ה- $null\ distribution$: $\chi^2\ distribution$, מסומן ב- $\chi^2(dx)$, כאשר df = דרגות החופש.

$\chi^2\ Table$:

- מראה אותו דבר כמו טבלת T (ההסתברות עד אינסוף).
- אבל כאן אי אפשר להניח סימטריות! צריך לבדוק לשני הערכים.

$\chi^2\ Test\ for\ goodness\ of\ fit$

בהינתן דטה-סט, שמניחים עליו התפלגות בדידה כלשהי, רוצים לבדוק האם הדטה מקיים את ההתפלגות שהנחנו. כלומר, H_0 אומרת שהדטה תואם להתפלגות הזו, ו- H_1 אומר שהדטה מתפלג אחרת.

נניח שנתונה pmf לא ידועה, בעזרת הטבלה הבאה:

Outcomes	ω_1	ω_2	$\dots$	ω_n
Probabilities	p_1	p_2	$\dots$	p_n

מניחים סט ערכים עבור ההסתברויות הללו (בד"כ נניח התפלגות ידועה כלשהי - בינומית, פואסון...).

המבחן צריך לקבוע האם סט ההסתברויות הללו הוא זה שיצר את ה- $data$ שנאסף.

<u>סימונים</u>: לכל תוצאה אפשרית ω_i:

$observed\ count\ O_i$ (כמה מה- $data\ points$ שייכות לקטגוריה i)

$expected\ counts\ E_i$, שחושבו מה- $null\ distribution\ probability\ table$ (כמה $data\ points$ מצפים שיהיו בקטגוריה i ע"פ ההתפלגות שמניחים. אם יש הסתברויות, נכפול אותן במספר הכולל של ה-data points).

ישנם 2 $test\ statistics$ שאפשר להשתמש בהם:

1. $Likelihood\ ratio\ statistic$

$$G = 2 \cdot \Sigma Q_i ln(\frac{Q_i}{E_i})$$

2. $Pearson's\ chi - squared\ statistic$

$$X^2 = \Sigma \frac{(Q_i - E_i)^2}{E_i}$$

את ה- $p - values$ מחשבים ע"פ הטבלה, וכך מחליטים האם לקבל או לדחות את H_0.

! דרגות החופש: $df = n - 1$ (כאשר n = מספר הקטגוריות/ התוצאות).

$\chi^2\ Test\, for\ homogeneity$

בהינתן מספר סטים של דטה שנדגם רנדומית, נרצה לדעת האם כל הסטים הללו נדגמו מאותה התפלגות.

כלומר, H_0 אומרת שכל הסטים נדגמו מאותה התפלגות (לא אומרים איזו), ו- H_A אומרת שלא.

ה- $Data$ שניתן כאן הוא מספר עבור כל תוצאה אפשרית לכל $data\ set$, כלומר לכל $data\ set\ i$ יש $observed\ count\ O_{i,j}$ לכל תוצאה אפשרית ω_j.

בעזרת ה- $null\ distribution$: $\chi^2(dx)$ אנו יכולים להעריך את ה- $expected\ counts$ לכל ה- $data\ sets$.

חישוב ה-test statistic הוא כמו למעלה, ואז קובעים את ה- $p - value$ ע"פ הטבלה ומחליטים האם לדחות את השערת האפס או לא.

! דרגות החופש: $(n-1)(m-1)$ כאשר n = מספר הקטגוריות, m = מספר הדוגמאות.

$F\ Test$

כמו $T\ Test$ רק עם n קבוצות של נתונים עם m $data\ points$ בכל אחת.

ה- $F\ distribution$ נגזרה מדטה נורמלי, בטווח $[0, \infty)$.

הנחה: הנקודה ה- j בקבוצה ה- i מתפלגת נורמלית עם תוחלת μ_i (של הקבוצה ה- i), ושונות כלשהי.

כלומר, $y_{i,j} \sim N(\mu_i, \sigma^2)$

- $H_0 = Null\ Hypothesis$: $\mu_1 = \mu_2 = ... = \mu_n$
- $test\ statistic$: $\frac{MS_B}{MS_W}$ כאשר: $MS_B = \frac{m}{n-1}\Sigma(\overline{y_i} - \overline{y})^2,\ MS_W = \frac{s_1{}^2 + s_2{}^2 + ... + s_n{}^2}{n}$

$$s_i^2 = \text{sample variance of group } i = \frac{1}{m-1}\sum_{j=1}^{m}(x_{i,j} - \bar{x}_i)^2.$$

-
- אם השערת האפס נכונה, אז ה- $test\ statistic$ הזה מתפלג לפי התפלגות $F_{n-1,\ n(m-1)}$.
- $Null\ Distribution$: $F\ statistic$ עם $n-1$ ו- $n(m-1)$ דרגות חופש.
- אם $\frac{MS_B}{MS_W} \sim F_{n-1,\ n(m-1)}$, אז H_0 נכונה. היחס אמור להיות קרוב ל-1.

$Percentile,\ Quantile$

- החציון (median, נקרא לפעמים Q2) הוא ה- $Percentile$ ה50.
  50% מהתוצאות נמצאות מתחתיו, ו-50% מעליו.
- ה-"first quantile" (או Q1) הוא ה- $Percentile$ ה25.

25% מהתוצאות נמצאות מתחתיו, ו-75% מעליו.

- ה-"third quantile" (או Q3) הוא ה-$Percentile$ ה75.
  75% מהתוצאות נמצאות מתחתיו, ו-25% מעליו.

<u>הערות:</u> Q2, בניגוד לחציון רגיל של קבוצה, לא צריך להיות איבר ששייך לקבוצה. באופן כללי, כל ה-quantiles לא צריכים להיות דווקא איברים בקבוצה. אם הנקודה הרצויה יוצאת בין שני מספרים, ניקח את האמצע ביניהם להיות התשובה.

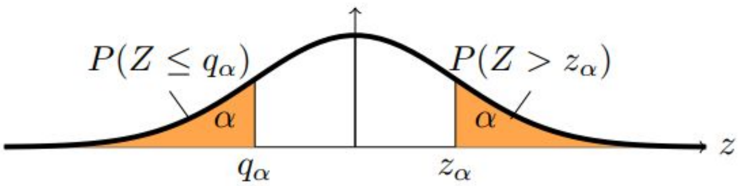
- הערכים הקריטיים משלימים ל-$quantiles$ כלומר הערך הקריטי שמתאים להסתברות p מסוימת הוא ה-$quantile$ ה-$1-p$. $c_p = q_{1-p}$.

$quantile$ של התפלגות הוא q_α עבורו: $P(X < q_\alpha) = \alpha$

- עבור התפלגות נורמלית נהוג לסמן את ה-$quantile$ ב-q_α.

$critical\ value$ של התפלגות הוא c_α עבורו: $P(X > c_\alpha) = \alpha$

- סימונים מיוחדים לערך קריטי: עבור $N(0,1)$ - z_α, עבור $t(n)$ - t_α.



$Confidence\ Interval$

המטרה היא לאמוד מה קורה באוכלוסייה כולה על פי תוצאה של מדגם. בד"כ רוצים לחשב ממוצע של משהו. הנתונים בשאלה: הממוצע שיצא במדגם, סטיית התקן באוכלוסייה ורמת הסמך הנדרשת ($confidence\ level$) אפשר לחשב סטייה מסוימת בין התוצאה שיצאה במדגם לבין התוצאה של כל האוכלוסייה. בסופו של דבר נאמר שהממוצע באוכלוסיה נע בין משהו למשהו.

<u>שלבי פתרון:</u>

1. מציירים פעמון, שמרכזו הוא התוצאה שיצאה במדגם (ממוצע המדגם).
2. לוקחים מרחק זהה לשני הכיוונים, והשטח בין שתי הנקודות הקיצוניות נקרא "רמת סמך", $= \alpha$.
3. השטח בקצוות של הפעמון הוא $\frac{\alpha}{2}$.
4. המרחק בין המרכז של הפעמון לנקודה השמאלית = המרחק לנקודה הימנית = ε. מוצאים אותו לפי הנוסחה הבאה: $\varepsilon = z \cdot \frac{\sigma}{\sqrt{n}}$
5. התשובה הסופית היא: $mean - \varepsilon < \mu < mean + \varepsilon$ (כאשר $mean$ = ממוצע המדגם).
   ! התשובה היא בבטחון של α (מה שנתנו לנו בתור רמת הסמך).

עבור דטה שמתפלג נורמלית $N(\mu, \sigma^2)$, עם $confidence\ level = 1 - \alpha$:

σ ידוע ← $z\ confidence\ interval\ for\ the\ mean$: $[\ \bar{x} - \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2}\cdot\sigma}{\sqrt{n}}\ ]$

המשמעות: אם נחזור על הניסוי המון פעמים, ב-$1-\alpha$ מהמקרים, μ יהיה באינטרוול שנקבל.

σ לא ידוע ← $t\ confidence\ interval\ for\ the\ mean$: $[\ \bar{x} - \frac{t_{\alpha/2}\cdot s}{\sqrt{n}},\ \bar{x} + \frac{t_{\alpha/2}\cdot s}{\sqrt{n}}\ ]$

כאשר s^2 הוא ה-$sample\ variance$ של הדטה, ומחושב ע"פ הנוסחה הבאה: $S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$

לשים 🖤 ה-α כאן הוא לא רמת הבטחון, אלא 1 פחות רמת הבטחון!

כאשר רוצים לקרב את σ^2 ← $\chi^2\ confidence\ interval\ for\ \sigma^2$: $[\ \frac{n-1}{c_{\alpha/2}}s^2,\ \frac{n-1}{c_{1-\alpha/2}}s^2\ ]$

- ל-t ול-χ^2 יש $n-1$ דרגות חופש.

- הכוונה ב-*margin of error* של p היא שהביטוי שמחברים לממוצע כדי לחשב את האינטרוול יהיה שווה ל-p הזה. ביטוי שקול הוא: לדעת את הפרמטר (למשל μ) "*within p*" ($p = 1\%$ למשל).

standard Bernoulli approximation אומר: $\sigma \leq 1/2$.

interval statistic: אינטרוול I_x המחושב מ-*data x*, הוא אינטרוול רנדומי, מכיוון ש-x כזה.

interval estimate: כאשר מעריכים את הערך של פרמטר.

Pivoting

יהי c קבוע, ו-$\bar{x}$ ממוצע הדגימות, ו-μ_0 *hypothesized mean* (לא ידועה).

אזי: $\bar{x}$ באינטרוול $\mu_0 \pm c$ ⇔ μ_0 באינטרוול $\bar{x} \pm c$

Polling

יהי *data* המתפלג ברנולי, כלומר: $x_1, ..., x_n \sim Ber(\theta)$ [θ לא ידועה, ורוצים להעריך אותה].

conservative normal $(1-\alpha)$ *confidence interval for* θ:

$$\left[\bar{x} - \frac{z_{\alpha/2}}{2\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right]$$

Large sample confidence interval

יהיו $x_1, ..., x_n$ משתנים מקריים בלתי תלויים שנדגמו מהתפלגות מסוימת בעלת תוחלת ושונות סופיים.

גרסה של משפט הגבול המרכזי אומרת שעבור n גדול: $\frac{\bar{x}-\mu}{s/\sqrt{n}} \approx N(0, 1)$

המשמעות: עבור n גדול מספיק, הביטוי הנ"ל מתפלג נורמלי-סטנדרטי.

עבור n גדול, ה-$(1-\alpha)$ *confidence interval* של μ הוא בקירוב:

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot z_{\alpha/2},\ \bar{x} + \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}\right]$$

Views of confidence intervals

View 1

הגדרת/ בניית CI עבור μ באמצעות סטנדרטיזציה של *point statistic*.

נתונים $x_1, ..., x_n \sim N(\mu, \sigma^2)$, כאשר σ ידוע.

$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, ולכן: $P(-z_{\alpha/2} < z < z_{\alpha/2} \mid \mu) = 1 - \alpha$

ניתן לעשות *pivot* ל: $P(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \mid \mu) = 1 - \alpha$

זה בעצם $(1-\alpha)$ CI: $x \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

גם t ו-χ^2 מתאימים לפרדיגמה הזאת: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

View 2

הגדרת/ בניית CI עבור μ באמצעות מבחני השערות. יש פרמטר θ לא ידוע, ו-*x test statistic*.

נוכל לקבוע H_0: כ: $\theta = \theta_0$ עם α *significant level*, ולבדוק האם x תומך בהשערה או לא.

- בהינתן x, $(1-\alpha)$ CI מכיל את כל ה-θ_0-ים עבורן H_0 לא נדחית.
- *type 1 error* של CI מתרחשת כאשר האינטרוול לא מכיל את הערך האמיתי של θ.

- עבור $(1-\alpha)\ CI$ אחוז המקרים שבהם תהיה טעות מסוג 1 של CI הוא α.

View 3

הגדרת CI ככל *interval statistic* בעל מאפיין מתמטי מסוים.

יהי x הנדגם מ- $f(x|\theta)$, עם פרמטר θ לא ידוע.

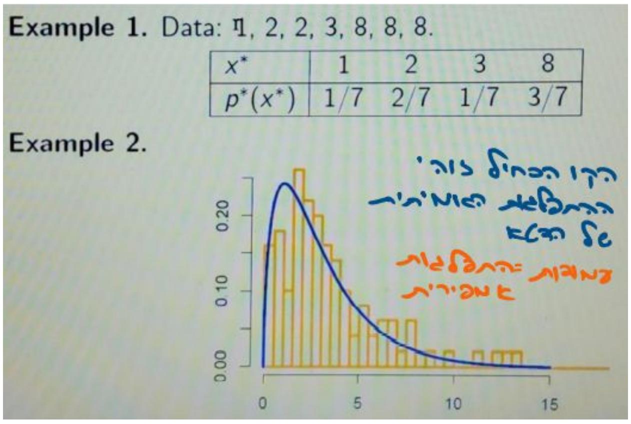
$(1-\alpha)\ CI$ של θ הוא *interval statistic* I_x כך ש- $P(I_x \text{ contains } \theta \mid \theta) = 1-\alpha$, עבור כל ערכי θ האפשריים, ולכן בפרט גם עבור הערך האמיתי של θ.

Bootstrapping

התפלגות אמפירית של ה- *data*:

יהיו $x_1, ..., x_n \sim F$.

נוכל לבדוק כמה פעמים הופיע כל ערך ולקבל התפלגות אמפירית על הדטה, שבקירוב שווה לאמיתית.



Resampling

בהינתן *sample* כלשהו, *resampling* לגודל m משמעותו לדגום באופן רנדומלי m דגימות עם חזרות מה- *sample* המקורי.

Resample probabilities = ההתפלגות האמפירית של כל ערך אפשרי מה- *sample*.

Bootstrap Resample הוא דגימת איברים מה- *sample* המקורי, עם חזרות, כאשר גודלו זהה תמיד לזה של ה- *sample* המקורי.

Empirical bootstrap

empirical bootstrap sample הוא *resample* באותו גודל n: $x^*_1, ..., x^*_n$. עקרון ה- *bootstrap* אומר שלקבוצה הזו יש התפלגות אמפירית F^*.

לכל *statistic* v שחושב מה- *sample sata* המקורי, נוכל להגדיר *statistic* v^* ע"י אותה נוסחה, רק שהוא יחושב מה- *resample data*.

The bootstrap setup is as follows:

1. $x_1, x_2, \ldots, x_n$ is a data sample drawn from a distribution F.
2. u is a statistic computed from the sample.
3. F^* is the empirical distribution of the data (the resampling distribution).
4. $x_1^*, x_2^*, \ldots, x_n^*$ is a resample of the data of the same size as the original sample
5. u^* is the statistic computed from the resample.

Then the bootstrap principle says that

1. $F^* \approx F$.

2. The variation of u is well-approximated by the variation of u^*.

Our real interest is in point 2: we can approximate the variation of u by that of u^*. We will exploit this to estimate the size of confidence intervals.

עקרון ה- *Bootstrap* עבור תוחלת:

בהינתן $x_1, ..., x_n \sim F$ עם תוחלת μ כלשהי, נסמן:

F^* = ההתפלגות האמפירית של הדטה (*resampling distribution*).

$x_1^*, ..., x_n^*$ = הדטה לאחר *Bootstrap Resample*.

ה- *Bootstrap Principle* אומר:

1) ההתפלגות של F^* זהה להתפלגות של F.

2) $\delta^* = \overline{x^*} - \overline{x} \approx \overline{x} - \mu = \delta$ (=הוריאציה של x).

3) $\overline{x} - \delta^*_{\alpha/2} \le \mu \le \overline{x} - \delta^*_{1-\alpha/2}$

<u>הערה:</u> העקרון עובד לכל סוג כלשהו של סטטיסטיקה, ולא רק לממוצע (תוחלת). הדבר היחיד שצריך לשנות הוא סוג ה- *statistic*.

מציאת confidence interval בעזרת עקרון ה- Bootstrapping

$\delta = \overline{x} - \mu$ מסמלת עד כמה ההתפלגות של $\overline{x}$ נמצאת מסביב לתוחלת μ.

לפי העקרון, נוכל לקרב את ההתפלגות של δ ע"י $\delta^* = \overline{x^*} - \overline{x}$, כאשר $\overline{x^*}$ הוא הממוצע של ה- *empirical bootstrap sample*.

לפי חוק המספרים הגדולים, את ההתפלגות של δ^* אפשר לקרב בעזרת מחשב שיסמלץ את ה- *resampling* מהדטה המקורי הרבה מאוד פעמים, ויחשב על כל *bootstrap sample* כזה את ה- *statistic* (למשל ממוצע).

לאחר מכן, נמיין את ה- δ^*ים שיצאו לנו בסדר עולה (יותר נכון לא יורד), מה שיביא לנו מעין *quantiles* של δ^*, שבעזרתם נוכל לקרב את ה- *quantiles* של δ.

דוגמה: כשמבקשים *confidence interval* של 80%, אז $\alpha = 0.2$, ולכן $\delta^*_{\alpha/2} = \delta^*_{0.1}$ ו- $\delta^*_{1-\alpha/2} = \delta^*_{0.9}$.

נגדיר את ה- *confidence interval* עבור μ כך: confidence interval $[\hat{\theta} - \delta^*_{1-\alpha/2}, \hat{\theta} - \delta^*_{\alpha/2}]$

Parametric bootstrap

ההבדל היחיד בין זה לבין ה- *empirical bootstrap* הוא המקור של ה- *bootstrap sample*. כאן אנחנו מחוללים את ה- *bootstrap sample* מהתפלגות פרמטרית.

מציאת *confidence interval* עבור פרמטר:

0. Data: $x_1, \ldots, x_n$ drawn from a distribution $F(\theta)$ with unknown parameter θ.
1. A statistic $\hat{\theta}$ that estimates θ.
2. Our bootstrap samples are drawn from $F(\hat{\theta})$.
3. For each bootstrap sample

$$x_1^*, \ldots, x_n^*$$

we compute $\hat{\theta}^*$ and the bootstrap difference $\delta^* = \hat{\theta}^* - \hat{\theta}$.

4. The bootstrap principle says that the distribution of δ^* approximates the distribution of $\delta = \hat{\theta} - \theta$.
5. Use the bootstrap differences to make a bootstrap confidence interval for θ.

confidence interval $[\hat{\theta} - \delta^*_{1-\alpha/2}, \hat{\theta} - \delta^*_{\alpha/2}]$

Linear Regression

הרעיון הוא למדל *data* דו-מקומי $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ לפונקציה + רעש, כלומר המודל הוא:
$y_i = f(x_i) + E_i$ (E_i נקרא "random error").
ההנחה היא שה-E_i-ים הם בלתי תלויים, עם שונות זהה - σ^2.
את הטעות של המודל ניתן למדוד באמצעות הנוסחה הבאה של *total squared error*:

$$\sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

המודל מאפשר לנו לחזות את הערך של y (שנקרא dependent or response variable), לכל ערך נתון של x (שנקרא independent or predictor variable).

ה-*Linear Regression* לינארי בפרמטרים $a, b, \ldots$, ומכאן שמו.
צריך לפתור אוסף של משוואות לינאריות, בהתאם למספר הנעלמים.

(1) *Simple Linear Regression*
רוצים למצוא את הקו (הלינארי) הפשוט ביותר המתאים ל-data. כלומר: $y_i = ax_i + b + E_i,\ E_i \sim N(0, \sigma^2)$
(ל-σ יש ערך קבוע שזהה עבור כל ה-data points).
המטרה היא למצוא את הערכים עבור a ו-b שנותנים את הקו המתאים ביותר, כלומר ממזערים את ה-total-squared error.
<u>הערה:</u> ממזערים את $\sum_{i=1}^{n} E_i^2$ ע"י מציאת הנגזרת (שימוש בנגזרות החלקיות) והשוואה ל-0.
רק עבור *Simple Linear Regression* יש נוסחאות בעזרתן ניתן לקצר את החישוב:

$$\overline{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \quad , \quad s_{xx} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$\overline{y} = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i \quad , \quad s_{xy} = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

נוסחאות עבור הפרמטרים האופטימליים: $\widehat{a} = \frac{s_{xy}}{s_{xx}} \quad , \quad \widehat{b} = \overline{y} - \widehat{a} \cdot \overline{x}$

(2) $Linear\ Regression$

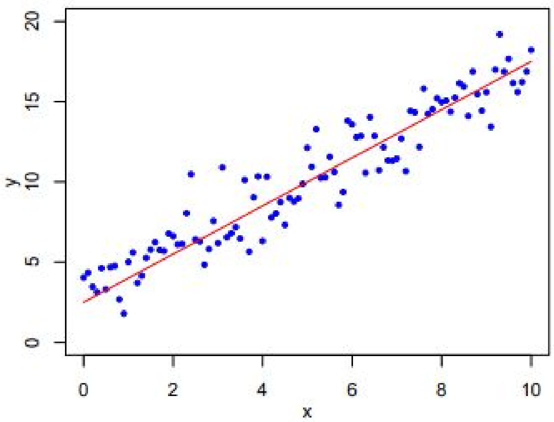
רוצים להתאים את הפולינום הטוב ביותר ל-data, למשל פולינום ממעלה שנייה - פרבולה. כלומר: $y_i = ax_i^2 + bx_i + c + E_i,\ E_i \sim N(0, \sigma^2)$ (גם כאן σ קבוע וזהה עבור כל ה-data points).

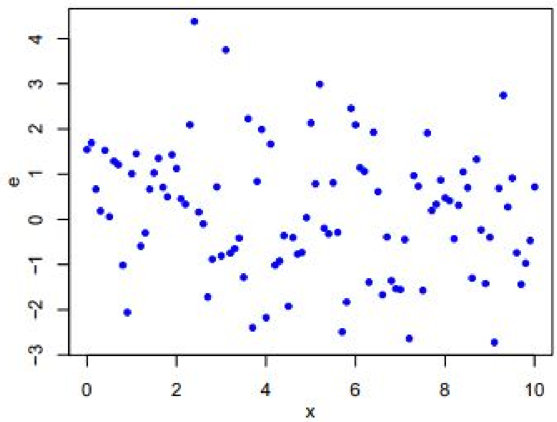
המטרה היא למצוא את הערכים עבור a, b, c שממזערים את ה-total squared error.

Homoscedastic

מושג המתייחס לפיזור אחיד של השגיאות סביב ה- $regression\ line$.

באיור השמאלי ניתן לראות את הקו הלינארי שהותאם ל-data, ובאיור הימני ניתן לראות את ה"שאריות" - המרווחים בין הנקודות לקו. ניתן לראות שהם מתפזרים באופן אחיד:





מדידת הטעות

נסמן את הערכים (האמיתיים) של y ב: $y = (y_1, \ldots, y_n)$

ואת הערכים שהותאמו ע"י המודל ב: $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_n)$

ישנן כמה דרכים לבטא את הטעות:

- $TSS = \sum (y_i - \bar{y})^2$ = total sum of squares = total variation.
- $RSS = \sum (y_i - \hat{y}_i)^2$ = residual sum of squares.
  RSS = unexplained by model squared error (due to random fluctuation)
- RSS/TSS = unexplained fraction of the total error.
- $R^2 = 1 - RSS/TSS$ is measure of goodness-of-fit
- R^2 is the fraction of the variance of y explained by the model.

Overfitting

אם נגדיל את הדרגה של הפולינום (הנבחר בתור ה-model function):

- הערך של R^2 יגדל
- הסיבוכיות של המודל תגדל

הדרגה האופטימלית היא $trade-off$ בין התאמה טובה לבין סיבוכיות המודל.

- אם כל ה-data points יהיו בדיוק על הקו, כלומר מעלת הפולינום = מספר הדגימות, אז $y = \hat{y}$ ו-$R^2 = 1$ אבל כמובן שאנו לא רוצים זאת - כי זה כנראה לא תואם בכלל למצב של פיזור ה-data במציאות. מסובך מידי.