התפלגויות (1)

	צפיפות	תוחלת	סטיית תקן (שורש שונות)
אחידה (רציפה) $X \sim U[a, b]$	$f(x) = \frac{1}{b-a}$ אם $a \le x \le b$	$\mu = \frac{b+a}{2}$	$\sigma = \frac{b-a}{2\sqrt{3}}$
נורמלית $X \sim N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ
בינומית $X \sim B(n, p)$	$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$	$\mu = np$	$\sigma = \sqrt{n(1-p)p}$
	“The probability of having n successes in k trials, with probability of success in one trial = p”		
פואסון (דיסקרטי) $X \sim P(\lambda)$	$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$ $, k \ge 0$	$\mu = \lambda$	$\sigma = \sqrt{\lambda}$
	“The probability of an event having k occurrences within a given interval T” Where λ is the mean amount of occurrences in a timespan of T. (θT where theta is the mean rate).		
גיאומטרית $X \sim G(p)$	$P(k) = (1-p)^{k-1} p$	$\mu = \frac{1}{p} - 1$	$\sigma = \frac{\sqrt{(1-p)}}{p}$
	"The Probability of needing k trials to get a success"		
אקספוננציאלית (רציף) $X \sim Exp(\theta)$	$f(t) = \theta \cdot e^{-\theta \cdot t}$	$\mu = \frac{1}{\theta}$	$\sigma = \frac{1}{\theta}$
	"The probability that t time passed until a success” Where θ is the rate of occurrences.		

Estimation / אומדנים (2)

משערכים פרמטרים של התפלגות כלשהי, בהינתן מדגם של האוכלוסייה.
תכונות של אומדנים:

Bias/הטייה	$B(\hat{\phi}) = E(\hat{\phi}) - \phi$	אומדן חסר הטיה מקיים $E(\hat{\phi}) = \phi$
MSE	$MSE(\hat{\phi}) = E((\hat{\phi} - \phi)^2) = V(\hat{\phi}) + B(\phi)^2$	אומדן חסר הטיה מקיים $MSE(\hat{\phi}) = V(\hat{\phi})$
Consistency/ Convergence in mean	$\lim_{n \to \infty} MSE(\hat{\phi}) = 0$	עבור מספר דגימות גדול האומדן שואף לפרמטר האמיתי

אומדן התוחלת: $\hat{\mu} = \frac{x_1+...+x_n}{n}$	זה אומדן חסר הטייה (נובע ממשפט הגבול המרכזי), וקונסיסטנטי (השונות קטנה כתלות ב-n וה-MSE שווה לה). סטיית התקן של האומדן היא $\frac{\sigma}{\sqrt{n}}$ כלומר האומדן משתפר ביחס ישר לגודל הדגימה וביחס הפוך לשונות של האוכלוסייה המקורית. (זה ה-**Standard Error of mean**)
אומדן השונות: $\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$	זה אומדן חסר הטייה. השונות של האוכלוסייה, אם יודעים את μ היא: $V = E((X - E(X))^2) = E(X^2) - E(X)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
אומדן קוריאנס: $\hat{\sigma}_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu}_X)(y_i - \hat{\mu}_Y)$	**אומדן קורלציה:** $\hat{r}_{XY} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X\hat{\sigma}_Y}$

(3) אומדני נראות מירבית / MLE

הגדרה: פונקציית הנראות (Likelihood) היא $L(\theta|X) = P(X|\theta)$ כאשר θ הם הפרמטרים של ההתפלגות.
הנראות היא דבר יחסי, אין לה ממש שימוש בפני עצמה.
לעיתים קרובות ממקסמים את לוג פונקציית הנראות. ניתן להשתמש ב**חוקי לוגים**:

$$log(x * y) = log(x) + log(y) \qquad log(x/y) = log(x) - log(y)$$

Maximum Likelihood Estimator, כשאנחנו רוצים לשערך את הפרמטר שנותן נראות מקסימלית של מודל.

1. מחשבים את פונקציית הנראות (בדרך כלל היא כמו פונקציית הצפיפות המתאימה, אבל הקלט שלה הוא הפרמטר). אם יש יותר מדגימה אחת, הנראות היא המכפלה של פונקציות הצפיפות שלהן.
2. מחשבים את לוג פונקציית הנראות.
3. גוזרים ומשווים ל-0 כדי למצוא נקודת מקסימום.

אפשר לחשב שונות של אומד אם לוקחים את הפתרון האנליטי.
דוגמאות:

מ"מ פואסון:	מ"מ בינומי:	מ"מ אקספוננציאלי (N אובסרווציות):
$L(\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$ $log(L) = l = klog(\lambda) - - log(k!) - \lambda$ נגזור ונשווה ל-0: $\frac{dl}{d\lambda} = \frac{k}{\lambda} - 1 = 0$ $\hat{\lambda} = k$	$L(p) = \binom{n}{k}p^k(1-p)^{n-k}$ $log(L) = l = log\binom{n}{k} + klogp + (n-k)log(1-p)$ נגזור ונשווה ל-0: $\frac{dl}{dp} = \frac{k}{p} - \frac{n-k}{1-p} = 0$ $\hat{p} = \frac{k}{n}$	$L(\theta) = \theta e^{-\theta t_1} \cdot ... \cdot \theta e^{-\theta t_N}$ $log(L) = l = N \cdot log(\theta) - \theta\sum_{i=1}^{N} t_i$ נגזור ונשווה ל-0: $\frac{dl}{d\theta} = \frac{N}{\theta} - \sum_{i=1}^{N} t_i = 0$ $\hat{\theta} = \frac{N}{\sum_{i=1}^{N} t_i}$

Maximum A Posteriori Estimator - MAP, לוקח בחשבון גם התפלגות פריורית על הפרמטר. למשל אולי אנחנו יודעים שהוא מתפלג נורמלי. אז:

$$Pr(\theta|X) = \frac{Pr(X|\theta)Pr(\theta)}{Pr(X)} \qquad log(Pr(\theta|X)) = log(Pr(X|\theta)) + log(Pr(\theta)) - log(Pr(X))$$

מאחר ו$log(Pr(X))$ לא תלוי ב-θ, כשנגזור לפי θ הוא יעלם. על כן:

$$\theta_{MAP} = arg\,max_{\theta}\, log(Pr(X|\theta)) + log(Pr(\theta))$$

Bootstrapping (4)

מוטיבציה: למדוד תכונות של אומדים, למשל סטיית התקן שלהם **(Standard Error)** (אומדים הם בעצמם משתנים מקריים).

1. בהינתן דגימות $x_1,..., x_n$, משתמשים בהם בתור התפלגות של האוכלוסייה.
2. עבור B איטרציות, בכל איטרציה, בוחרים n דגימות (עם החזרה), ומחשבים את ה-Estimator לפיהן. (למשל עבור אומדן התוחלת עושים את הממוצע של הדגימות).
3. כעת יש לנו B אומדנים שונים ואפשר לאמוד את הממוצע ואת השונות/סטיית התקן שלהם:

$$\bar{\theta} = \frac{\hat{\theta}_1 + ... + \hat{\theta}_B}{B} \qquad\qquad \hat{\sigma}^2_{Boot} = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\theta})^2$$

גם $\hat{\sigma}_{Boot}$ הוא משתנה מקרי וגם לו יש שונות. אפשר להראות שמתקיים:

$$V(\hat{\sigma}_{Boot}) = \frac{1}{n^2} + \frac{1}{B \cdot n}$$

את שני האיברים אפשר להקטין על ידי הגדלת גודל המדגם, ואת האיבר השני אפשר להקטין גם על ידי הוספת איטרציות בוטסטראפ.

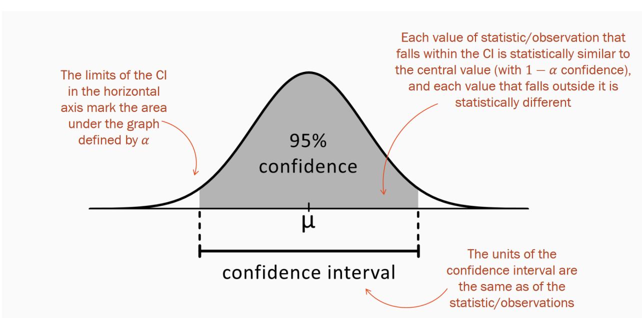
מציאת אחוזונים: האחוזון ה-α הוא הערך כך ש-$\alpha\%$ מהדגימות קטנות שווה לו.

1. נניח שיש לנו B תוצאות אומדני בוטסטראפ. **נמיין** אותן מהקטן לגדול $\hat{\theta}_1,..., \hat{\theta}_B$
2. האחוזון ה-x הוא ערך הבוטסטראפ באינדקס $i = (B + 1) \cdot x$, כלומר $\hat{\theta}_i$.
3. אם i אינו מספר שלם, ניקח את הממוצע של עיגול למעלה ולמטה: $\frac{\hat{\theta}_{\lfloor i \rfloor} + \hat{\theta}_{\lceil i \rceil}}{2}$

תיקון Bias בעזרת בוטסטראפ: אם $\hat{\theta}$ זה האומד המקורי ו-$\bar{\theta}$ זה ממוצע הבוטסטראפ, אז ניתן לתקן את ההטייה שלו כך:

$$\hat{\theta}_{BC} = \hat{\theta} - Bias(\hat{\theta}) = \hat{\theta} - (\bar{\theta} - \hat{\theta}) = 2\hat{\theta} - \bar{\theta}$$

חישוב Confidence Interval - CI בעזרת בוטסטראפ: נניח שיש לנו סטטיסטי מסוים שחישבנו מדגימה, למשל ממוצע, ואנחנו רוצים לחשב טווח סביב התוצאה שקיבלנו, כך שהערכים הנופלים בטווח הזה סטטיסטית דומים לערך שקיבלנו. (בציור, ציר ה-x זה ערכים של הסטטיסטי, ציר ה-y זה כמה הוא סביר).
אז נקבע α כלשהו. השטח הלבן בציור זה אלפא. וכל מה שנופל בשטח האפור אנחנו בטוחים בו ב-$1 - \alpha$ בטחון.



Normal CI: מניחים שהדאטא מגיע מהתפלגות נורמלית.

$$\hat{\theta} = \hat{\theta} \pm Z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}_{Boot}$$

כאשר $Z_{1-\frac{\alpha}{2}}$ הוא z-score שנמצא בטבלה.

Percentile CI:

נחשב בוטסטראפים, הגבולות הם: $U = \hat{\theta}^{b}_{1-\frac{\alpha}{2}}$ $L = \hat{\theta}^{b}_{\frac{\alpha}{2}}$

(כאשר זה מסמן את האחוזונים מדגימות הבוטסטראפ).

Pivot CI: זו השיטה הכי טובה. הגבולות הם: $U = 2\hat{\theta} - \hat{\theta}^{b}_{\frac{\alpha}{2}}$ $L = 2\hat{\theta} - \hat{\theta}^{b}_{1-\frac{\alpha}{2}}$

כאשר האיבר הראשון זה החישוב המקורי של האומדן והאיברים השניים זה אחוזונים מחישובי הבוטסטראפ. זה לאו דווקא סימטרי סביב הערך המרכזי.

Jackknifing: כמו בוטסטראפ, אבל עושים n איטרציות ובכל פעם מחשבים בלי דגימה אחת.

$$\sigma^2_{jack} = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \bar{\theta})^2$$

וכך מתקנים bias בעזרת jackknifing:

$$\hat{\theta}_{BC} = n\hat{\theta} - (n-1)\bar{\theta}$$

(5) בדיקות השערות / Hypothesis Testing

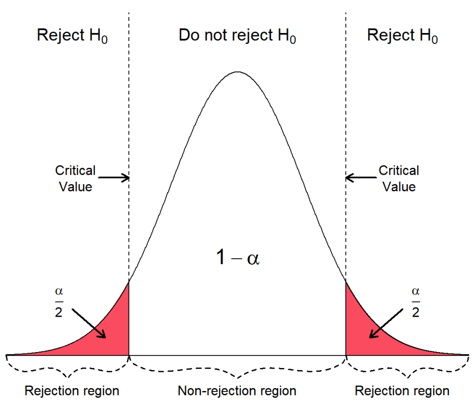
אנו רוצים להחליט איזו מבין שתי השערות סבירה יותר. H_0 היא ה-null hypothesis, בדרך כלל הפשוטה יותר, ו-H_1 היא האלטרנטיבית. (הן משלימות).

סוגי טעות ומדדים:

1. **טעות מסוג 1:** $Pr(reject\ H_0|H_0) = \alpha$ זה נקרא גם **False alarm rate** וגם **Significance**.
2. **טעות מסוג 2:** $Pr(not\ reject\ H_0|\bar{H}_0) = \beta$

$(1-\beta)$ נקרא **עוצמה**. ככל שהעוצמה גדולה יותר יש פחות שגיאה מסוג 2.

	Accept H_0	Reject H_0
H_0 is True	(TN)	(FP) Error type I
H_0 is False	(FN) Error type II	TP



P-Value: הגדרה: בהינתן הסטטיסטי (על ציר ה-x) זה ה-α הקטן ביותר שיאפשר לשלול את H_0. ככל שהוא קטן יותר, אנחנו יותר בטוחים בשלילת השערת ה-0.

T-test: בהינתן שני מדגמים מהתפלגויות **נורמליות** שונות $x_1,..., x_n$ ו-$y_1,..., y_n$, אלה ההשערות שלנו:

$H_1: \mu_x \neq \mu_y\ \ or\ \ \mu_x > \mu_y\ \ or\ \ \mu_x < \mu_y$ $H_0: \mu_x = \mu_y\ \ or\ \ \mu_x \leq \mu_y\ \ or\ \ \mu_x \geq \mu_y$

כלומר המבחן בודק אם הממוצעים שווים או שונים. אפשר גם להשוות מול קבוצה אחת לממוצע ידוע של

אוכלוסייה. המבחן מבוסס על התפלגות t - כמו נורמלית עם זנבות יותר עבים. היא מקבלת פרמטר dof וככל שהוא שואף לאינסוף התפלגות t שואפת להתפלגות נורמלית סטנדרטית.

1. נבחר רמת סיגניפיקנטיות α.
2. נחשב דרגות חופש. $dof = n - 1$ (עבור קבוצה מול אוכלוסיה) $dof = n - 2$ (עבור שתי קבוצות)
3. נחשב את הסטטיסטי: $t = \frac{\bar{x}-\mu_0}{\sigma_x/\sqrt{n}}$ (קבוצה מול אוכלוסייה).
4. נבדוק בטבלת t מה הערך הקריטי לפי ה-dof וה-α והאם ההשערה חד-צדדית או דו-צדדית. אם הסטטיסטי קיצוני נדחה את השערת ה-0, אחרת נקבל אותה.

ANOVA: להשוואה של $k > 2$ קבוצות אשר באות מהתפלגויות נורמליות.

$H_0: \mu_1 =...= \mu_k$ $\qquad$ $H_1: Otherwise$

(6) טרנספורמציות ובדיקת נורמליות

בדיקת נורמליות (Shapiro-Wilk Test) - כשיש לנו מדגמים נבדוק אם הם מתפלגים נורמלית. במקרה זה נשתמש במבחנים שמניחים נורמליות (t-test,ANOVA).

1. בהינתן דגימות $x_1,..., x_n$ נגדיר: $H_0: x_i \sim N$ $\qquad$ $H_1: otherwise$
2. נמיין את הדגימות. מעתה נניח $x_1 \leq ... \leq x_n$. נגדיר $m = \frac{n}{2}$ אם n זוגי) ו-$m = \frac{n-1}{2}$ אם n א"ז).
3. נחשב את הסטטיסטי: $W = \frac{\left(\sum_{i=1}^{m} a_i(x_{n+1-i} - x_i)\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ (כאשר a_i הם קבועים בטבלה והם תלויים ב-n)
4. נשלול את H_0 אם $W_{calculated} < W_\alpha$ (כאשר α היא הסיגניפיקנטיות שבחרנו ו-W_α בטבלה (וגם תלוי ב-n))

טרנספורמציה - לעיתים ניתן לקבל דאטא שמתפלג נורמלי על ידי הפעלת טרנספורמציה לא ליניארית על הדאטא המקורי, למשל לוג, אקספוננט העלאה בריבוע.

הסרת אאוטליירים בשיטת IQR - לעיתים ייתכן שמספר קטן של אאוטליירים גרמו לדאטא להיראות כאילו הוא לא מתפלג נורמלי.

1. נגדיר Q_1 הרבעון הראשון (השברון ה-0.25) ונגדיר Q_3 הרבעון השלישי (השברון ה-0.75). (נחשב אותם כמו בבוטסטראפ). נגדיר: $IQR = Q_3 - Q_1$
2. נגדיר אאוטליירים להיות הדגימות שמקיימות:

$$sample < Q_1 - \frac{3}{2}IQR \qquad sample > Q_3 + \frac{3}{2}IQR$$

כלומר נשמור רק את הדגימות בתחום $[Q_1 - \frac{3}{2}IQR, Q_3 + \frac{3}{2}IQR]$

(7) מבחנים א-פרמטריים

נועדים לשימוש כשהתנאים למבחנים פרמטרים לא מתפלגים (ההתפלגות לא נורמלית) או שיש מעט דגימות.
הם פחות חזקים ממבחנים פרמטריים ולרוב מסתמכים על דירוג.

דירוג: דוגמא להלן:

value	10	30	50	50	50	70	90	90	100
rank	1	2	4	4	4	6	7.5	7.5	9

Mann-Whitney U-test: מבחן א-פרמטרי, בעל העוצמה הכי גבוהה מבין המבחנים הא-פרמטריים. מקביל א-פרמטרי של t-test, אבל נועד להשוואה של התפלגויות ולא של ממוצעים, כלומר:

$$H_0: X\ is\ distributed\ like\ Y \qquad H_1: otherwise$$

1. בהינתן דגימות $x_1,..., x_{n_x}$ ו- $y_1,..., y_{n_y}$, נדרג אותן ביחד, תוך זכירה של אילו דגימות הן מ-x ואילו מ-y.
2. נחשב את R_x, סכום הדרגות של דגימות ה-x ואת R_y, סכום הדרגות של דגימות ה-y.
3. נחשב את הסטטיסטים (נשים לב שקל לחשב אחד מהשני):

$$U = n_x n_y + \frac{n_x(n_x+1)}{2} - R_x \qquad U' = n_x n_y + \frac{n_y(n_y+1)}{2} - R_y = n_x n_y - U$$

4. נמצא בטבלה את הערך הקריטי U_{α,n_x,n_y} (נשים לב שצריך $\alpha(2)$ כלומר דו צדדי בטבלה). נשלול את H_0 אם $Max(U, U') \geq U_{\alpha,n_x,n_y}$.

<u>הערה:</u> U-test מניח כי השונות של שתי ההתפלגויות דומה, אחרת, ייתכן שנקבל את H_0 עבור שתי התפלגויות שונות. ניתן לפרמל את השערת ה-0 גם כך: $H_0: median_x = median_y$
ולכן לרוב הוא משמש להשוואה בין חציונים.
<u>עוצמה:</u> כאשר הדאטא נורמלי העוצמה היא 95.5% משל t-test. בכל התפלגות של הדאטא, העוצמה היא יותר מ86.4% של t-test.

Mood's-test: מקרה מיוחד של **χ^2 Test For Independence**. בניגוד ל-U-test, לא מניח שהשונות של ההתפלגויות זהה. הוא גם פחות עוצמתי מ-U-test ולכן עדיף להשתמש ב-U-test כשאפשר.

$$H_0: median_x = median_y \qquad H_1: otherwise$$

1. בהינתן דגימות $x_1,..., x_{n_x}$ ו- $y_1,..., y_{n_y}$, נמיין (לא נדרג) אותן **ביחד**. ונמצא את החציון המשותף.
2. לכל קבוצה, נספור כמה ערכים מעל החציון וכמה מתחת לחציון **בפועל (observed)**. ונשים בטבלה. אלה ה-Observed.
3. לכל קבוצה, נחשב כמה ערכים מעל החציון וכמה מתחת לחציון **צפויים (expected)**. ונשים בטבלה.
4. נחשב את הטבלה $\frac{(observed-expected)^2}{expected}$ וניקח את סכום הערכים בה. זה הסטטיסטי שלנו **χ^2**.
5. נגדיר $DOF = (l-1)(k-1) = (2-1)*(2-1) = 1$. ונמצא בטבלה את הערך הקריטי $\chi^2_{\alpha,DOF}$ (בטבלה שראינו בתרגול 0.05 זה בעצם 0.95).
6. אם $\chi^2 \geq \chi^2_{critical}$ נשלול את H_0.

observed	A	B	
מעל החציון			R1
מתחת לחציון			R2
	C1	C2	T

expected	A	B
מעל החציון	(R1*C1)/T	(R1*C2)/T
מתחת לחציון	(R2*C1)/T	(R2*C2)/T

<u>עוצמה:</u> עבור דאטא המתפלג נורמלי, העוצמה 64% משל t-test ו-67% משל U-test. על כן כדאי להשתמש רק כשהשונות של ההתפלגויות שונה.

Test For Independence χ^2: מכליל את Mood's-test. הוא בודק אם יש קורלציה בין שני משתנים מקריים, כאשר אנחנו מקבלים דאטא שמשתייך לקטגוריות במשתנים המקריים. במבחן mood, המשתנים הם הקבוצה X או Y, והיות הדגימה מעל או מתחת לחציון המשותף עם הקטגוריות 1. מעל 2. מתחת. כמות הקטגוריות של המשתנה המקרי הראשון היא k, וכמות הקטגוריות של המשתנה המקרי השני היא l, מעבר לכך, החישוב זהה (עם טבלאות גדולות יותר ו $DOF = (l-1)(k-1)$).

$H_0: X \text{ and } Y \text{ are independent} \qquad H_1: \text{not so}$

Kruskal-Wallis: המקביל הא-פרמטרי ל-ANOVA.

$H_0: \text{The distributions of all groups is the same} \qquad H_1: \text{otherwise}$

1. נמיין את הדאטא מכל הקבוצות ביחד, **ונדרג** אותם ביחד (ונזכור מאיזו קבוצה כל אחד).
2. נחשב את הסטטיסטי: $H = \frac{12}{N(N+1)}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right) - 3(N+1)$
   כאשר: k הוא מספר הקבוצות, n_i היא כמות התצפיות בקבוצה i, R_i הוא סכום הדרגות של קבוצה i, N הוא כמות התצפיות הכוללת.
3. אם יש תצפיות עם ערכים זהים (תיקו בדירוג) צריך לחשב ערך מתקן: $C = 1 - \frac{\sum_{i=1}^{m}(t_i^3 - t_i)}{N^3 - N}$
   כאשר: m זו כמות הקבוצות של תצפיות בתיקו, t_i זו כמות התצפיות בתיקו בקבוצה i.
4. הסטטיסטי הסופי הוא $H_C = \frac{H}{C}$.
5. כאשר ($k = 3$ וגם $n_i \geq 5$ לכל i) או ($k > 3$ וגם $n_i \geq 4$ לכל i) נגדיר $dof = k - 1$ וניתן להשתמש ב $\chi^2_{\alpha,dof}$ בתור אומדן לערך הקריטי H_α. אחרת יש טבלה ספציפית לזה.
6. אם $H_C \geq H_{critical}$ נשלול את השערת ה-0.

Multiple Comparison / השוואה מרובה (8)

כשעושים כמה מבחנים עם significance שהוא α, ההסתברות שעשינו טעות מסוג 1 (False Positive) גדלה. ספציפית היא $\alpha^* = 1 - (1-\alpha)^n \geq \alpha$.
נסמן ב-α_0 את האלפא שאנו רוצים בסך הכול. נרצה למצוא α (למבחן בודד) הכי גדול שאפשר שעדיין יאפשר ל-α_0 להיות קטן.

Šidák Correction - שיטה זו מניחה שהמבחנים הבודדים בלתי תלויים אחד בשני והיא מבטיחה כי $\alpha^* = \alpha_0$:

$$\alpha = 1 - (1-\alpha_0)^{\frac{1}{n}}$$

Bonferroni Correction - שיטה זו אינה מניחה כי המבחנים בלתי תלויים אחד בשני והיא מבטיחה כי $\alpha^* \leq \alpha_0$. כשהם בלתי תלויים, שיטה זו יותר גרועה משיטת sidak. ו-α קטן מהר כתלות ב-n.

$$\alpha = \frac{\alpha_0}{n}$$

FDR - Benjamini-Hochberg procedure - השיטות האחרות הסתמכו על הרעיון שכל המבחנים הבודדים צריכים להיות צודקים בו זמנית. כאן אנחנו מוכנים לסבול קצת False Discovery Rate. (כלומר לשלול את H_0 כאשר H_0 נכון).

1. נניח שיש לנו n היפותזות $H_{(1)},..., H_{(n)}$. עשינו n מבחנים וקיבלנו p-values $p_1,..., p_n$.
2. נגדיר $0 \leq q \leq 1$, שהוא ה-False Discovery Rate שאנו מאפשרים.
3. נמיין את ה-p-values מהקטן לגדול. מעתה נניח $p_1 \leq ... \leq p_n$.
4. עבור כל p-value, נבדוק אם $p_i \leq \frac{i}{n} \cdot \frac{q}{c(n)}$. נסמן ב-$k$ את ה-i הגדול ביותר עבורו זה מתקיים.
5. נשלול את השערת ה-0 עבור כל ההיפותזות $H_1,..., H_{(k)}$.
6. ניתן לחשב גם $q_{corrected}$, שהם כמו p-values מתוקנים כך שהם מעידים עבור הליך ה-FDR מה ה-α_0 המינימלי שיאפשר לשלול את H_0 עבורם. $q_{corrected(i)} = \frac{p_i \cdot n \cdot c(n)}{i}$.
7. לבסוף נחשב את ה- adjusted q^* כך שערכי ה-q יהיו מונוטוניים ונדווח עליהם.

$ordered\ p$	i	$q_i = \frac{i \cdot q}{n}$	$significant?$	$q_{corrected}$	q^*
0.0003	1	0.0025	True	0.006	0.006
0.0015	2	0.0050	True	0.015	0.015
0.0036	3	0.0075	True	0.024	0.024
0.0065	4	0.0100	True	0.032	0.026
0.0065	5	0.0125	True	0.026	0.026
0.0154	6	0.0150	False		
0.0211	7	0.0175	False		

<u>הערה:</u> ברוב המוחלט של המקרים, כאשר המבחנים הם בלי תלות או בעלי תלות חיובית, הוכח כי $c(n) = 1$ הוא מספיק טוב.
רק כאשר המבחנים הם בעלי תלות שלילית כדאי להשתמש ב- $c(n) = \sum_{i=1}^{n} \frac{1}{i}$

(9) מבחני פרמוטציות

	sample 1	sample 2	...	sample n
gene 1	x_{11}	x_{12}		x_{1n}
gene 2	x_{21}	x_{22}		x_{2n}
⋮				
gene k	x_{k1}	x_{k2}		x_{kn}

	sample 1	sample 2	...	sample n
label	healthy	sick		sick

אנו מניחים שהנתונים שלנו נראים כמו בטבלה. כלומר יש דגימות עם כמה קטגוריות, ולכל אחת יש לייבל.
נניח שיש לנו סטטיסטי של הטבלה שמתאר קשר בין הלייבלים לשאר הדאטא. $T = f(X, L)$ <u>לדוגמא:</u> ההפרש בין הביטוי של גנים בחולים לעומת בריאים.
המטרה: לראות איך הסטטיסטי מתפלג תחת H_0.

1. נגדיר את H_0 (בדרך כלל שאין קשר, בדוגמא לעיל זה שהסטטיסטי הוא 0). ואת H_1.
2. נבצע R פרמוטציות. כלומר ניקח את הטבלה, ונסדר אקראית את הלייבלים הקיימים (בלי החזרה). בכל פרמוטציה נחשב את הסטטיסטי T. זה ייתן לנו התפלגות אמפירית על הסטטיסטי תחת H_0.
3. נחשב את ה-p-value של הסטטיסטי המקורי על פי ההתפלגות האמפירית.

חישוב p-value: אם מניחים/ידוע שהסטטיסטי T מתפלג נורמלי תחת השערת ה-0, אז אפשר לאמוד:

$$\hat{\mu} = \frac{T_1+...+T_R}{R} \qquad \hat{\sigma}^2 = \frac{1}{R-1}\sum_{i=1}^{R}(T_i - \hat{\mu})^2$$

ואז לחשב p-value תחת ההנחה כי $T \sim N(\hat{\mu}, \hat{\sigma}^2)$. כלומר לחשב Z-score כך: $Z = \frac{T-\hat{\mu}}{\hat{\sigma}}$ ואז לראות בטבלה לפי ה-Z מה ה-p-value.
<u>דרך כללית:</u> לספור את כמות הפרמוטציות בעלות ערך **יותר קיצוני** מתוך כלל הפרמוטציות.
לדוגמא, עבור השערה חד צדדית בצד שמאל ($H_1: T < x$): $p = \frac{\#(T_i \le T)}{R}$.
עבור השערה דו צדדית ($H_1: T \neq x$): $p = \frac{\#(|T_i| \ge |T|)}{R}$

בחירת כמות פרמוטציות R: אם מחשבים p-value לפי ספירה, בחירת כמות הפרמוטציות משפיעה על הרזולוציה של ה-p-value. לדוגמא עבור $R = 1,000$ הערכים האפשריים ל-p הם $0, 0.001, 0.002...$.
אם באמת מקבלים 0, לא מדווחים על 0 אלא כותבים: $p < \frac{1}{R}$.

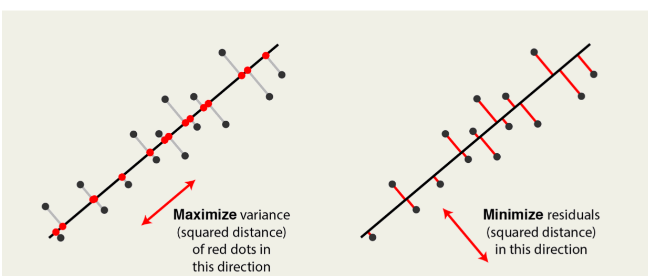
פרמוטציות על נתונים בזוגות: אם רוצים לעשות מבחן פרמוטציות על נתונים בזוגות, למשל לפני ואחרי, מחליפים אקראית בין הלפני והאחרי עבור כל אינדיבידואל.

(10) הורדת מימד

נניח שיש לנו n תצפיות, כל אחת עם d פיצ'רים. אז נתאר אותם בתור מטריצה $D_{d\times n}$. אנו רוצים להקטין את המימד, כלומר להוריד את כמות הפיצ'רים ובכל זאת לשמור על שונות.
עבור כל השיטות:
1. תמיד **נמרכז** את הדאטא. עבור כל שורה (פיצ'ר) נחשב את הממוצע, ונחסיר את הממוצע מכל הערכים בשורה. כעת הממוצע של כל שורה הוא 0.
2. כשיש לנו פיצ'רים שנמדדים ביחידות שונות, נרצה לבצע **סטנדרטיזציה** כך שלכל פיצ'ר תהיה השפעה זהה. עבור כל שורה נחשב את סטיית התקן (שורש אומד השונות), ונחלק את הערכים בשורה בו.

$$D = \begin{array}{cc} \text{Samples (n)} & \\ \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{d1} & \cdots & d_{dn} \end{pmatrix} & \text{Features (d)} \end{array} \rightarrow \quad X = \begin{pmatrix} \frac{d_{11}-m_1}{\sigma_1} & \cdots & \frac{d_{1n}-m_1}{\sigma_1} \\ \vdots & \ddots & \vdots \\ \frac{d_{d1}-m_d}{\sigma_d} & \cdots & \frac{d_{dn}-m_d}{\sigma_d} \end{pmatrix}$$

PCA - השיטה הבסיסית ביותר. היא Unsupervised Learning. מוצאים את תתי המרחבים ממימד 1 אשר שומרים על השונות הכי גבוהה. זו שיטה ליניארית.
1. נבצע את העיבוד הבסיסי לעיל.
2. השיטה לא מאפשרת datapoints חסרים. אם עבור sample מסוים חסרים הרבה פיצ'רים, אפשר לזרוק את ה-sample. אם חסרים קצת פיצ'רים, אפשר לשים בהם את הממוצע של הפיצ'ר המתאים.
3. נמצא את הווקטורים העצמיים והערכים העצמיים של המטריצה X^TX. נסמנם $u_1,\dots,u_d$. ו-$\lambda_1 \geq \dots \geq \lambda_d$.
4. $PC1$ הוא u_1, $PC2$ הוא u_2 וכו. **השונות המוסברת (Explained Variance)** של PC_i היא: $\frac{\lambda_i}{\lambda_1+\dots+\lambda_d}$
5. לאחר מכן אולי נטיל את X על u בשביל ויזואליזציה כך: X^Tu.



<u>הערה:</u> בתמונה פירושים אחרים של PCA:
<u>הערה:</u> איך בוחרים כמה PCs? עושים Scree plot, ציר ה-x זה המספר של ה-PC, ציר ה-y זה השונות המוסברת, וזה נראה כמו מרפק את עוצרים בפינה של המרפק.
<u>הערה:</u> PCA **משמר יותר מבנה גלובאלי מאשר לוקאלי**. השונות היא אופרטור ריבועי ולכן: א. נותנים יותר משקל לשונות של נקודות רחוקות (מבנה גלובלי). ב. לאאוטליירים יש השפעה חזקה.

FDA/LDA - דומה ל-PCA. גם שיטה ליניארית. בניגוד אליו, זו בעיה Supervised כלומר יש לייבלים והדאטא מחולק למחלקות. כאן $DV1$ זה הציר שמשמר את השונות הכי גבוהה בין לייבלים שונים, אבל מקטין את השונות בתוך כל לייבל.
כלומר פותרים את הבעיה $S_Bu = \lambda S_Wu$. (כאשר S_B, S_W כפי שהוגדרו ב-ANOVA).
אם כמות הקבוצות היא g, אז ניתן לקבל $g-1$ $DV's$.

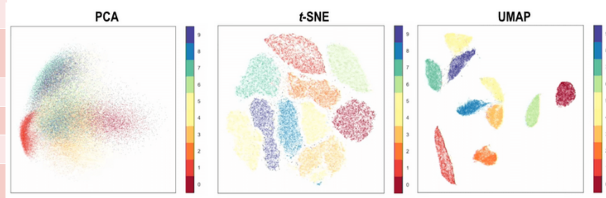
tSNE - בניגוד ל-PCA, שיטה זו מנסה **לשמור על מבנה לוקאלי ולא גלובלי**. פותרת בעיות של SNE למשל בעיית ה-Crowding שנקודות שנפרסו על הרבה נפח במרחב הגדול צריכות לתפוס מעט נפח במרחב הקטן.

1. מודדים דמיון בין כל זוג נקודות במרחב הגדול.
2. שמים באופן אקראי את הנקודות במרחב הקטן. מבצעים gradient descent. באופן איטרטיבי, משנים את המיקום של הנקודות במרחב הקטן, כך שלכל שתי נקודות אם הן היו קרובות במרחב הגדול, הן יהיו קרובות במרחב הקטן. נסמן ב-$p_{i|j}$ את הדמיון בין הנקודות במרחב הגדול ואת $q_{i|j}$ את הדמיון במרחב הקטן.

בעיות: שומר רק על מבנה לוקאלי, יקר חישובית, ומעשית ניתן לעשות embed רק למרחב דו או תלת מימדי.

UMAP - מתבסס על עקרונות טופולוגיים. דומה ל-tSNE, משמר גם מידע גלובאלי וגם לוקאלי! אפשר לבחור לאיזה מימד להוריד, על כמה שכנים להתבסס בחישוב, מה המרחק המינימלי במרחב הקטן, כמה איטרציות לעשות.

	PCA	t-SNE	UMAP
Technique family	Matrix factorization	Neighbor graph	Neighbor graph
(Primarily) preserved data structure	Global	local	Local+global
Will you get the same result each time?	Yes	No (stochastic)	Yes
Learning algorithm?	No	Yes	Yes
Linear?	Yes	No	No
Efficient?	Yes	No	Yes
Interpretable?	Yes (axes have actual meaning)	No	No

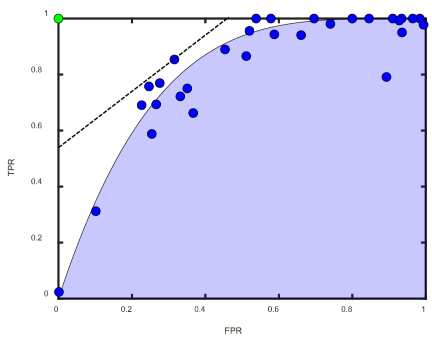


(11) קלסיפיקציה ורגרסיה לוגיסטית

הערכה של מסווג - יש כמה מדדים מוכרים.

Error Rate	Accuracy = 1-Error	דיוק, Precision, PPV	רגישות, Sensitivity, True Positive Rate, TPR	ספציפיות, Specificity, True Negative Rate	False Positive Rate
$\frac{FN+FP}{total}$	$\frac{TP+TN}{total}$	$\frac{TP}{TP+FP}$	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{FP}{TN+FP}$

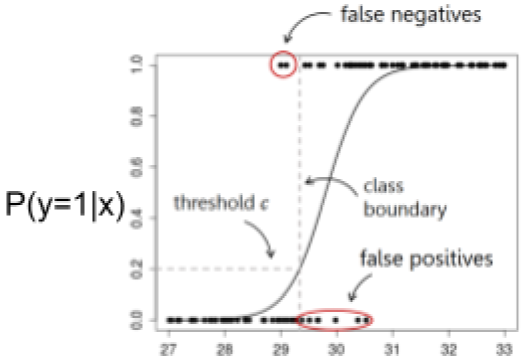
Receiver operating curve (ROC): כדי לבחור פרמטרים למודל. כל נקודה היא עבור פרמטרים אחרים למסווג. ציר ה-X זה FPR, ציר ה-Y זה TPR. נבחר את הפרמטרים שמתאימים לנקודה שממקסמת את $TPR - FPR$. (הכי קרובה לפינה השמאלית למעלה).



	predicted true	predicted false	
really true	TP	FN	T
really false	FP	TN	F
	P	N	

רגרסיה לוגיסטית (Logistic Regression): נניח שיש לנו תצפיות $x_1,...,x_n$ ולייבלים $y_1,...,y_n \in \{0, 1\}$. נרצה למדל את ההסתברות שתצפית חדשה שייכת ללייבל כלשהו. כלומר, את $log(\frac{Pr(y=1|x)}{Pr(y=0|x)}) = log(\frac{Pr(y=1|x)}{1-Pr(y=1|x)}) = logit$ = **log-odds**. אם זה גדול מסף כלשהו t נסווג את x בתור 1, ואחרת נסווג אותו בתור 0.

1. נרצה למצוא ביטוי ליניארי: $log(\frac{Pr(y=1|x)}{1-Pr(y=y|x)}) = w_0 + x_{(1)}w_1 +...+ x_{(d)}w_d$.
2. זה יתבצע על ידי מודל Maximum likelihood. פונקציית הלוס היא
$$l = \sum_{i=1}^{n} y_i log(\frac{e^{w\cdot x_i}}{1+e^{w\cdot x_i}}) + (1-y_i)log(\frac{1}{1+e^{w\cdot x_i}}) =$$
$$= \sum_{i=1}^{n} y_i log(\sigma(w,x_i)) + (1-y_i)log(1-\sigma(w,x_i))$$
3. נגזור אותה ונשתמש ב-Gradient Descent כדי למצוא את ה-w שממזער אותה.
4. נמצא את הסף t באמצעות ROC למשל.
5. לאחר מכן בסיווג ניתן גם לחשב את $Pr(y=1|x)$ במפורש בהינתן $log(\frac{Pr(y=1|x)}{1-Pr(y=y|x)})$ כך:
$$Pr(y = 1|x) = \frac{e^{wx}}{1+e^{wx}}$$



<u>הערה:</u> בגרף הבא, את c ניתן לחשב מהסף t כמו בסעיף 5, כאשר $t = wx$.