פונקצית הסתברות של מ"מ:
$0 \le p(X=x) \le 1$

ניסויים ב"ת והתפלגויות בדידות:

ניסוי ברנולי Ber(p): ניסוי בעל שתי תוצאות האפשריות: 'הצלחה'- S ו'כישלון'- F :

$bin\ /\ n$ $\quad \boxed{Var(X)=pq} \quad \boxed{EX=p}$.

התפלגות בינומית Bim(n,p): ההסתברות לקבל בדיוק k הצלחות בסדרה של n ניסויים שלכל אחד הסתברות הצלחה p . $\boxed{P(k)=\binom{n}{k}p^k q^{n-k}}$

סכום בינומים: אם $X \sim B(n,p)$ וגם $Y \sim B(m,p)$ (אותו p בשניהם) אז : $X+Y \sim B(m+n,p)$

התפלגות גיאומטרית Geo(p):
ההסתברות להצלחה ראשונה בנסיון ה- k כאשר ההסתברות ההצלחה בכל ניסוי היא p . $\boxed{p(k)=q^{k-1}p}$

חוסר זיכרון: אם ידוע שלא הייתה הצלחה עד לניסוי ה-m אז ההסתברות כי היא תקרה בניסוי m+k הינה עדיין כמו בנוסחה.

התפלגות בינומית שלילית NB(m,p):
ההסתברות לקבל את ההצלחה ה- m -ית בדיוק בניסיון ה- k כאשר לכל ניסוי הסת' הצלחה p .

$\boxed{p(k)=\binom{k-1}{m-1}p^m q^{k-m}}$ כמו סכום של גיאומטרי

התפלגות היפר-גיאומטרית HG(N,D,n): בכד נמצאים N כדורים שחורים ו- M כדורים לבנים. מוציאים n כדורים באקראי ללא החזרה. ההסתברות שבין הכדורים שהוצאו נמצאים k כדורים שחורים.

$$\boxed{P(k)=\frac{\binom{D}{k}\binom{N-D}{n-k}}{\binom{N}{n}} \quad n<N,\ D<N}$$

מרובה סוגים: כאשר $n_1+..+n_k=n$ ובוחרים n מ N.

$$P(n_j \text{ of type } 1\le j\le k)=\frac{\binom{N_1}{n_1}\binom{N_2}{n_2}\cdot\ldots\cdot\binom{N_k}{n_k}}{\binom{N}{n}}$$

קירוב בינומי להיפר גיאומטרי: כאשר N גדול מאוד ביחס ל- n , אזי תוצאת החישוב עם ההחזרה (הבינומית) קרובה לתוצאת החישוב בלי החזרה (ההיפר גיאומטרית). $p=D/N,\ n=n$.

$$\boxed{X \sim HG(n,N,M) \Rightarrow X \sim Bin\left(n,\frac{M}{N+M}\right)}$$

התפלגות פואסון בדידה Pois(λ): מספר האירועים ביחידת זמן נתונה, אם ידוע כי הם מתרחשים בקצב ממוצע קבוע $\lambda>0$ ובאופן ב"ת בפרק הזמן מאז האירוע האחרון. $\boxed{p(k)=\frac{e^{-\lambda}\lambda^k}{k!}}$

מירוץ פואסון: $\boxed{X_1|X_1+X_2=n \sim Bin\left(n,\lambda_1/(\lambda_1+\lambda_2)\right)}$

קירוב פואסוני לבינומי: כאשר n גדול ו-p קטן אז ניתן לקרב את הבינומי לפואסוני עם $\lambda=np$.

הערה! ניתן לעשות קירוב נורמלי לבינומי-מג"מ!

סכום פואסונים ב"ת: $X \sim Pois(\lambda)$, $Y \sim Pois(\mu)$
$(X+Y) \sim Pois(\lambda+\mu)$.

פיצול פואסון: אם $Z \sim Pois(\mu)$ ואז מתרחש אחד משני מאורעות בהסתברויות $p_a+p_b=1$ אז נקבל מ"מים ב"ת שמתפלגים $Pois(p_a\cdot\mu), Pois(p_b\cdot\mu)$.

ב"ת!: מתקיים פיצול לכל שתי אפשרויות של "האנשים" שמגיעים. אלו שני פואסון ב"ת!

התפלגות אחידה Uni(k,n): כל ערך מתוך טווח ערכים מסוים, מתקבל באותה הסתברות.

סטטיסטי מספיק- דוגמא:

$$f_X=\prod\frac{1}{\mu}e^{-\frac{1}{\mu}(x_i-\theta)}=\left(\frac{1}{\mu}\right)^n e^{-\frac{1}{\mu}\sum x_i}e^{-\frac{n\theta}{\mu}}$$

ס"מ ל- μ : $S=\sum X_i$. ס"מ ל- θ :

מוסיפים $I\{X_{(1)}>\theta\}$ $S=X_{(1)}$. ס"מ ל- (θ,μ)

$S=\left(X_{(1)},\sum X_i\right)$ ($h(\bar{x})=1$).

דוגמא-2:

$$f_X=\prod \beta x_i^{\beta-1}=\beta^n\left(\prod x_i\right)^{\beta-1}$$

$$S=\prod x_i \text{ or } \sum \ln x_i$$

$$\boxed{\frac{1}{n}\sum X_i^2-\bar{X}^2=\frac{1}{n}\sum(X_i-\bar{X})^2}$$

$$\boxed{\sum(X_i-\theta)^2=\sum(X_i-\bar{X})^2+\sum(\bar{X}-\theta)^2}$$

התפלגות משותפת ושולית:

התפלגות משותפת: חיתוך המאורעות
$\boxed{P(X=x,Y=y)=P_{XY}(x,y)}$

התפלגות שולית: $\boxed{P_X(x)=\sum_y P_{XY}(x,y)}$

אי-תלות מ"מ בדידים:
מ"מ הם ב"ת אם"ם
$\boxed{\forall x,y:\ P_{XY}(x,y)=P_X(x)\cdot P_Y(y)}$

אם הסתברות משותפת היא מכפלת שתי פונקציות, האחת של x והשנייה של y, $f(x)\cdot g(y)$ אזי ניתן להסיק שקיימת אי-תלות בין המשתנים.

הערה: פונקציות ותת-קבוצות זרות של מ"מ בלתי תלויים הן בלתי תלויות.

תוחלת: $\boxed{E(X)=\sum_{x\in R_X} xP(X=x)}$

($\sum_x |x|\cdot P(X=x)<\infty$

$E(X+Y)=EX+EY \quad E(aX+b)=aEX+b$

$E(g(X,Y))=\sum_x\sum_y g(x,y)\cdot P(X=x,Y=y)$

$E(g(X))=\sum_x g(x)\cdot P(X=x)$

כלל הכפל: אם X,Y ב"ת אז $\boxed{E(XY)=E(X)\cdot E(Y)}$
זוג מ"מ שמקיים את הנוסחה: **בלתי מתואמים.**

לכל שני מ"מ: $\boxed{E(XY)=\sum_x\sum_y xyP(X=x,Y=y)}$

חוק המספרים הגדולים: עבור X_i ב"ת ו-

החלש: לכל $EX_i=\mu$ אז: $\boxed{\frac{1}{n}\sum_{i=1}^n X_i\to\mu}$

החזק: $P(|\bar{X}_n-\mu|\ge\varepsilon)\to 0 \quad \varepsilon>0$

$P\left(\lim_{x\to\infty}|X_n-\mu|=0\right)=1$

נוסחת הזנב: $\boxed{E(X)=\sum_{k=1}^{\infty}P(X\ge k)}$ ($X\in\mathbb{N}$)

תוחלת מותנית:

$\boxed{E(X|Y=y)=\sum_x x\cdot P(X=x|Y=y)}$,

$\boxed{E(Y)=\sum_x E(Y|X=x)\cdot P(X=x)}$

אם X,Y ב"ת אז $E(X|Y)=E(X)$

במקרה הרציף: $E(X|Y=y)=\int_{-\infty}^{\infty}xf_{X|Y}(x|y)dx$

נוסחת ההחלקה:
$\boxed{E(X)=\sum_y E(X|Y=y)\cdot P(Y=y)}$

אם $E(X|Y)=g(Y)$
אז $\boxed{E(X)=E(E(X|Y))=E(g(Y))}$

במקרה הרציף

$$E(Y)=\int_{-\infty}^{\infty}f_x(x)\left[\int_{-\infty}^{\infty}y\cdot f_{Y|X}(y|x)dy\right]dx=\int_{-\infty}^{\infty}f_x(x)\cdot E(Y|X=x)dx$$

נוסחאות נוספות:

$E(X+Y|Y)=E(X|Y)+Y$

$E(X+Y|Y=y)=E(X|Y=y)+y$

$E(XY|Y=y)=y\cdot E(X|Y=y) \quad E(XY|Y)=Y\cdot E(X|Y)$

$E(X^2)=E(E(X^2|Y=y)) \qquad E(XY)=E(E(XY|Y=y))$

$E(X^2|Y=y)=Var(X|Y=y)+(E(X|Y=y))^2$

משפט wald: X_i ב"ת $EX_i=\mu$. N מ"מ ב"ת בכל ה- X_i , $S_N=X_1+\cdots+X_N$

$\boxed{E(S_N)=EX\cdot EN=\mu\cdot EN}$

QQ-PLOT:

i	1	…	n
i/n+1			
$\xi_{i/n+1}$			
תצפיות	$X_{(1)}$		$X_{(n)}$

ציר x: שברונית תיאורטיים- $\xi_{i/n+1}$

ציר y: תצפיות ממוינות לפי הסדר.

דוגמא:

מוצאים כיצד לבטא ξ_p ע"י $p=i/n+1$

$F_X(x)=x^\beta \quad F_X(\xi_p)=P(X\le\xi_p)=p$

$\xi_p^{\ \beta}=p \xrightarrow{\beta=2} \xi_p=\sqrt{p}$

קו ישר בגרף- הנחת ההתפלגות נכונה.

שונות וסטיית תקן:

$\boxed{Var(X)=E(X^2)-(E(X))^2=E((X-E(X))^2)}$

$Var(aX+b)=a^2\cdot Var(X)$

$\boxed{Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)}$

אם X,Y **ב"ת:** $Var(X+Y)=Var(X)+Var(Y)$ וגם $Var(X|Y)=Var(X)$. c קבוע אמ"מ $Var(c)=0$.

$\boxed{SD(X)=\sqrt{Var(X)}} \quad SD(aX+b)=|a|\cdot SD(X)$

שונות מותנית: $Var(Y|X)=E\left((Y-E(Y|X))^2|X\right)$

$\boxed{Var(Y)=E\left(Var(Y|X)\right)+Var\left(E(Y|X)\right)}$

קו-ואריאנס: $Cov(X,X)=Var(X)$

$\boxed{Cov(X,Y)=E\left((X-\mu_x)(Y-\mu_Y)\right)=E(XY)-E(X)E(Y)}$

$\boxed{Cov(X,Y)=\frac{1}{2}(Var(X+Y)-Var(X)-Var(Y))}$

1) אם X ו- Y ב"ת אז $Cov(X,Y)=0$.
בד"כ ההיפך לא נכון $Cov(X,Y)=0 \not\Rightarrow X,Y$ ב"ת.
2) אם $Cov(X,Y)=0$ אז X,Y בלתי מתואמים.
ב"ת $\Leftarrow$ בלתי מתואמים. (ההפך לא בהכרח נכון).
3) $Cov(X,Y)=Cov(Y,X)$ - סימטריות.
4) $Cov(aX+b,cY+d)=acCov(X,Y)$.
5) $Cov(X,Y+Z)=Cov(Y,X)+Cov(X,Z)$.
6) לכל a קבוע $Cov(X,a)=0$.

תקנון: $X^*=\frac{X-\mu_X}{\sigma_X}$, כאשר: $Corr(X,Y)=E\left(X^*Y^*\right)$

אינטגרלים: $\int_0^\infty \lambda\cdot e^{-\lambda x}dx=1$, $\int_0^\infty e^{-z^2}dz=\sqrt{\pi}$

$\int_{-\infty}^{\infty}x\cdot e^{-\frac{x^2}{2}}dx=0 \qquad \int_0^\infty S^{n-1}\cdot e^{-S}dS=(n-1)!$

התפלגויות רציפות:

פונקצית הצפיפות:

$\boxed{P\left(X\in[a,b]\right)=P(a\le X\le b)=\int_a^b f_x(x)dx}$

$\boxed{F_X(x)=P(X\le a)=\int_{-\infty}^a f_x(x)dx}$

$P(a\le X\le b)=P(X\le b)-P(X\le a)$

אם $f_x(x)$ פונקצית צפיפות של X מ"מ רציף אז $\int_{-\infty}^{\infty}f_x(x)dx=1$ ותמיד $f_x(x)\ge 0$.

הסת' שלמה: $\boxed{f_X(x)=\sum_k f_{X|Y}(x|Y=k)P(Y=k)}$

תוחלת של רציף: $\boxed{E(X)=\int_{-\infty}^{\infty}x\cdot f_x(x)dx}$.

$\boxed{E(g(X))=\int_{-\infty}^{\infty}g(x)\cdot f_x(x)dx}$

התוחלת קיימת רק כאשר $\int_{-\infty}^{\infty}|x\cdot f_x(x)|dx<\infty$.

נוסחת הזנב $\boxed{E(X)=\int_0^\infty(1-F_x(x))dx=\int_0^\infty P(X>x)dx}$

לשים-לב: באזורים בהם $F_X(x)=0$, אז $\int 1$.

שונות של רציף:

$\boxed{Var(X)=\int_{-\infty}^{\infty}x^2\cdot f_X(x)dx-\left(\int_{-\infty}^{\infty}x\cdot f_X(x)dx\right)^2}$

התפלגות	ש.מומנטים	אנ"מ
		$\hat{\mu}=\bar{X}$
$N(\mu,\sigma^2)$		$\hat{\sigma}^2=\frac{1}{n}\sum(X_i-\bar{X})^2$
$Pois(\lambda)$		$\hat{\lambda}=\bar{X}$
$U(0,\theta)$	$\tilde{\theta}=2\bar{X}$	$\hat{\theta}=X_{(n)}$
$\exp(1/\theta)$		$\hat{\theta}=\bar{X}$
$\exp(\lambda)$		$\hat{\lambda}=1/\bar{X}$
$Bin(n,p)$		$\hat{p}=\bar{X}$
$Ber(p)$		$p=\bar{X}$

סטטיסטי מספיק:

$\theta=(\mu,\sigma^2)$ ס"מ ל- θ $\left(\sum X_i^2,\sum X_i\right)$

ס"מ לשונות כשהתוחלת ידועה: $\sum(X_i-\mu)^2$

ס"מ לתוחלת כשהשונות ידועה: $\bar{X}$

$Ber(p)$, $\exp(\lambda)$, $Pois(\lambda)$ - $\sum X_i$

פונקצית התפלגות מצטברת:

הגדרה: $\boxed{F_x(x)=P(X\le x)}$.

בדיד: $\boxed{F_x(x)=\sum_{t\le x}P(X=t)}$ **רציף:** $\boxed{F_x(x)=\int_{-\infty}^x f_x(x)dx}$

1) $0\le F_x(x)\le 1$. 2) $F_x(x)$ לא יורדת.

3) $\lim_{x\to-\infty}F_x(x)=0$ $\quad \lim_{x\to\infty}F_x(x)=1$.

4) $F_x(x)$ רציפה מימין. 5) $f_x(x)=\frac{d}{dx}F_x(x)$.

6) $F_X(x) \sim Uni[0,1]$ לכל פונקציה.

אם $E(X)<\infty$ אז $\lim_{x\to\infty}x\cdot(1-F_x(x))=0$.

התפלגות נורמלית: שימוש בטבלה עבור מ"מ

נורמלי $Z\sim N(0,1)$ $\boxed{f_Z(z)=\frac{1}{\sqrt{2\Pi}}e^{-z^2/2}}$

$F_X\to P(Z\le a)=\Phi(a) \qquad P(a\le Z\le b)=\Phi(b)-\Phi(a)$

סכום נורמלים: $X\sim N(\mu_1,\sigma_1^2)$ ו- $Y\sim N(\mu_2,\sigma_2^2)$
ב"ת אז $\boxed{X+Y\sim N(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)}$.

מנת נורמלים: מתפלגת קושי. **נורמלי ב-^{2}:** גאמא

משפט הגבול המרכזי: $X_1,X_2,\ldots,X_n$ ב"ת, ש"ה,

$E(X_i)=\mu$, $Var(X_i)=\sigma^2$ **נגדיר:** $\bar{X}_n=\frac{1}{n}\sum_{i=1}^n X_i$

אז: $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\underset{n\to\infty}{\Rightarrow}N(0,1)$ כלומר $\boxed{\bar{X}_n\sim N\left(\mu,\frac{\sigma^2}{n}\right)}$

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

$S_n=\sum_{i=1}^n X_i$ $\boxed{N(n\cdot\mu,n\sigma^2)}$ $\boxed{\frac{S_n-n\mu}{\sqrt{n}\sigma}\to N(0,1)}$

התפלגות אחידה (יוניפורמית): $U[0,1]\to f_X(x)=1$

$E(X)=\frac{a+b}{2} \qquad F_X\to P(c\le X\le d)=\frac{d-c}{b-a}$

$\boxed{f_x(x)=\begin{cases}\frac{1}{b-a} & a\le x\le b\\ 0 & otherwise\end{cases}} \qquad Var(X)=\frac{(b-a)^2}{12}$

סטטיסטי הסדר: אם $X_1,X_2,\ldots,X_n\sim U[0,1]$ ב"ת:

$\boxed{(\min|\max=x)\sim Uni[0,x]}$ $f_{\min}(x)=n(1-x)^{n-1}$ $f_{\max}(x)=nx^{n-1}$

$E(X_{\min})=\frac{b}{n+1} \qquad E(X_{\max})=\frac{n}{n+1}$

סטנדרטיזציה: $X\sim U[a,b]\Leftrightarrow\frac{X-a}{b-a}\sim U[0,1]$

מעבר לאקספוננט: $\left(-\frac{\ln(1-U)}{\lambda}\right)\sim\exp(\lambda)$, $U[0,1]$

התפלגות אקספוננציאלית: זמן בין מאורעות פואסון.

$Var(X)=1/\lambda^2 \quad E(X)=1/\lambda \quad P(X>a)=e^{-\lambda a}$

$\boxed{f_X(x)=\begin{cases}\lambda e^{-\lambda x} & x\ge 0\\ 0 & otherwise\end{cases}} \quad P(a\le X\le b)=e^{-\lambda a}-e^{-\lambda b}$

מינימום של אקספוננט הוא אקספוננט:
$X\sim Exp(\lambda),Y\sim Exp(\mu)\Rightarrow Min(X,Y)\sim Exp(\lambda+\mu)$

תחרות בין אקספוננציאלים: X,Y ב"ת

$X\sim Exp(\lambda),Y\sim Exp(\mu)\Rightarrow\boxed{P(X<Y)=\frac{\lambda}{\lambda+\mu}}$

סכום עם גיאומטרי: $X_i\sim\exp(\lambda)$, $N\sim Geo(p)$

אז: $S_N=\sum_{i-1}^N X_i\sim\exp(\lambda p)$

תכונת חוסר זיכרון: אם נתון כי עד רגע t לא היה מופיע, אז הזמן עד למופע הבא מתחיל מהתחלה

$\boxed{P(T>t+s|T>t)=P(T>s)}$ $E(X-Y|X>Y)=EX$.

משפט: נתון מ"מ T , נגדיר $G(t)=P(T>t)$. אם $G(t+s)=G(t)\cdot G(s)$, קיים λ כך ש $T\sim Exp(\lambda)$.

התפלגות Gamma: $X_1,X_2,\ldots,X_n$ מ"מ ב"ת,

$X_i\sim Exp(\lambda)$, נגדיר $Y=\sum_{i=1}^n X_i$, $Y\sim Gamma(n,\lambda)$.

$$\boxed{f_X(x)=\begin{cases}\frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{\Gamma(n)}=\frac{\lambda e^{-\lambda x}(\lambda x)^{n-1}}{(n-1)!} & x\ge 0\\ 0 & otherwise\end{cases}}$$

$\boxed{Gamma(1,\lambda)=\exp(\lambda)}$ $\boxed{\Gamma(n)=(n-1)\Gamma(n-1)}$ $\Gamma\left(\frac{1}{2}\right)=\sqrt{\Pi}$

כאשר n שלם: $\Gamma(n)=(n-1)!$

חיבור Gamma: אם $X\sim Gamma(r,\lambda)$,
$Y\sim Gamma(s,\lambda)$ ב"ת אז $\boxed{X+Y\sim Gamma(r+s,\lambda)}$.

חילוק Gamma: עבור X,Y כנ"ל יתקיים
$\frac{X}{X+Y}\sim Beta(r,s)$. המנה והסכום בשני הנ"ל ב"ת.

$x\sim\Gamma(n,\lambda),\ ax\sim\Gamma(n,\lambda/a)$

רווח סמך:

ר"ס לתוחלת במדגם מהתפלגות נורמאלית:

כאשר σ ידוע:

$$Q=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\sim N(0,1) \qquad \mu=\bar{X}\pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

כאשר σ לא ידוע:

$$Q=\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t_{(n-1)} \qquad \mu\in\bar{X}\pm t_{(n-1),1-\alpha/2}\frac{S}{\sqrt{n}}$$

תכונות:
1. זהו רווח סמך הקצר ביותר ברמת סמך נתונה.
2. אורכו קבוע מראש רק כאשר השונות ידועה.
3. אורך הרווח עולה עם הגדלת רמת הסמך והשונות.
4. אורך הרווח קטן עם עליה בגודל המדגם

גודל המדגם הדרוש כדי שבביטחון של $1-\alpha$ הסטייה בין ממוצע המדגם ל μ לא תעלה על d :

$$2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\le 2d\to \boxed{n\ge\frac{z_{1-\alpha/2}^2\sigma^2}{d^2}}$$

ר"ס ל- θ במדגם מהתפלגות אחידה:

$$Q\left(X_{(n)},\theta\right)=\frac{X_{(n)}}{\theta} \qquad \theta\in\left[X_{(n)},\frac{X_{(n)}}{\alpha^{1/n}}\right]$$

ר"ס לפרופורציה במדגם מהתפלגות נורמאלית:

$$X_i\sim Ber(p),\ \hat{p}=\bar{X}=\frac{1}{n}\sum X_i=\frac{X}{n} \qquad X\sim Bin(n,p)$$

$$Q=\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}\sim N(0,1) \qquad p\in\hat{p}\pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

גודל המדגם כדי שבביטחון של $1-\alpha$ הסטייה בין הפרופורציה במדגם ל p לא תעלה על d :

$$n\ge\frac{z_{1-\alpha/2}^2\hat{p}(1-\hat{p})}{d^2},\ p(1-p)\le 0.25,\ \boxed{n\ge\frac{z_{1-\alpha/2}^2}{4d^2}}$$

ר"ס לשונות במדגם מהתפלגות נורמאלית:

כאשר μ ידוע:

$$Q=\frac{1}{\sigma^2}\sum(X_i-\mu)^2\sim\chi^2_{(n)}$$

$$\sigma^2\in\left[\frac{\sum(X_i-\mu)^2}{\chi^2_{(n),1-\alpha/2}},\frac{\sum(X_i-\mu)^2}{\chi^2_{(n),\alpha/2}}\right]$$

כאשר μ לא ידוע:

$$Q=\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{(n-1)} \qquad S^2=\frac{1}{n-1}\sum(X_i-\bar{X})^2$$

$$\sigma^2\in\left[\frac{(n-1)S^2}{\chi^2_{(n-1),1-\alpha/2}},\frac{(n-1)S^2}{\chi^2_{(n-1),\alpha/2}}\right]$$

תכונות:
1. אינו רווח סמך אופטימאלי עבור רמת סמך נתונה.
2. רווח סמך עבור סטיית תקן: שורש על שני הקצוות.

ר"ס ל- α במדגם מהתפלגות $F_X(x)=x^{\alpha}$ (Gamma)

$$\hat{\alpha}=\frac{n}{-\sum\ln X_i} \qquad Y=-\ln(X),\ Y\sim\exp(\alpha),$$

$$Q=2\alpha\sum Y_i\sim Gamma\left(n,\frac{1}{2}\right)=\chi^2_{(2n)},\ f_X=\alpha x^{\alpha-1}$$

$$\alpha\in\left[\frac{\chi^2_{(2n),\alpha/2}}{2\sum Y_i},\frac{\chi^2_{(2n),1-\alpha/2}}{2\sum Y_i}\right] \qquad \sum Y_i\sim Gamma(n,\alpha)$$

ר"ס להפרש תוחלות בשני מדגמים ב"ת:

כאשר σ ידוע:

$$Q=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}}\sim N(0,1)$$

$$\mu_X-\mu_Y\in(\bar{X}-\bar{Y})\pm z_{1-\alpha/2}\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}$$

כאשר σ לא ידוע והשונויות זהות:

$$S_P^2=\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}$$

$$Q=\frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{S_P^2\left(\frac{1}{n}+\frac{1}{m}\right)}}\sim t_{(n+m-2)}$$

$$\mu_X-\mu_Y\in(\bar{X}-\bar{Y})\pm t_{(n+m-2),1-\alpha/2}\sqrt{S_P^2\left(\frac{1}{n}+\frac{1}{m}\right)}$$

כאשר σ לא ידוע והשונויות שונות:

$$\mu_X-\mu_Y\in(\bar{X}-\bar{Y})\pm t_{(df),1-\alpha/2}\sqrt{\left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)}$$

$$df=\left(\frac{S_X^2}{n}+\frac{S_Y^2}{m}\right)^2\Big/\frac{S_X^4}{n^2(n-1)}+\frac{S_Y^4}{m^2(m-1)}$$

ר"ס להפרש תוחלות בשני מדגמים מזווגים:

בונים מדגם של הפרשים

$$\mu_X-\mu_Y=\mu_D \qquad D_i\sim N(\mu_D,\sigma_D^2) \qquad \bar{D}=\frac{1}{n}\sum D_i$$

$$S_D^2=\frac{1}{n-1}\sum(D_i-\bar{D})^2 \qquad \mu_D\in D\pm t_{(n-1),1-\alpha/2}\frac{S_D}{\sqrt{n}}$$

ר"ס ליחס שונויות בשני מדגמים ב"ת:

$$Q=\frac{S_X^2}{S_Y^2}\frac{\sigma_Y^2}{\sigma_X^2}=\frac{S_X^2}{S_Y^2}\frac{1}{\lambda}\sim F_{(n-1,m-1)}$$

$$\lambda=\frac{\sigma_X^2}{\sigma_Y^2}\in\left[\frac{S_X^2}{S_Y^2}F^{\alpha/2}_{(m-1,n-1)},\frac{S_X^2}{S_Y^2}F^{1-\alpha/2}_{(m-1,n-1)}\right]$$

ר"ס להפרש פרופורציות בשני מדגמים ב"ת:

$$Q=\frac{(\hat{p}_X-\hat{p}_Y)-(p_X-p_Y)}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}} \qquad p_x-p_y\in$$

$$(\hat{p}_X-\hat{p}_Y)\pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}+\frac{\hat{p}_Y(1-\hat{p}_Y)}{m}}$$

התפלגויות דגימה:

התפלגות נורמאלית:

$$X_1,\cdots,X_n\sim N(\mu,\sigma^2)$$

$$\boxed{X_{0.95}=Z_{0.95}\cdot\sigma+\mu}\ \boxed{Z_{0.15}=-Z_{0.85}}$$

התפלגות הממוצע:

$$\bar{X}\sim N(\mu,\sigma^2/n)\to\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}\sim N(0,1)$$

הסתברות שהסטייה המוחלטת לכל היותר d :

$$P(|\bar{X}-\mu|\le d)=2\Phi\left(\frac{d}{\sigma/\sqrt{n}}\right)-1$$

גודל המדגם כדי ש: $P(|\bar{X}-\mu|\le d)\ge 1-\alpha$ -

$$n\ge\left(Z_{1-\alpha/2}\frac{\sigma}{d}\right)^2$$

חוק התקנון: $\mu-\sigma,\underbrace{\sigma-2\sigma}_{14\%},\underbrace{2\sigma-3\sigma}_{2\%}$

34% 14% 2%

התפלגות חי-בריבוע:

תכונות:
1. צפיפות לא סימטרית עם זנב ימני.
2. התפלגות חיובית.
3. עבור n>50, קרוב לנורמאלי.
4. אדטיביות: $\chi^2_{(n)}+\chi^2_{(m)}=\chi^2_{(n+m)}$

קשר ל-Gamma ולנורמאלי:

$$\boxed{\begin{array}{l}x\sim\Gamma(n,\lambda)\\ 2\lambda x\sim\Gamma(n,1/2)\end{array}}$$

$$W_n=\sum Z_i^2\sim\chi^2_{(n)}\overset{D}{=}Gamma\left(\frac{n}{2},\frac{1}{2}\right)$$

$$W_n=\frac{1}{\sigma^2}\sum(X_i-\mu)^2\sim\chi^2_{(n)}$$

$$\frac{W_n}{n}\xrightarrow{P}1 \qquad \frac{W_n-n}{\sqrt{2n}}\xrightarrow{D}N(0,1)$$

$$E(W_n)=n \qquad Var(W_n)=2n$$

התפלגות שונות המדגם:

תוחלת ידועה: $\hat{\sigma}^2=\frac{1}{n}\sum(X_i-\mu)^2$

לא ידועה: $\hat{\sigma}^2=\frac{1}{n}\sum(X_i-\bar{X})^2$

$$\boxed{\frac{\sum(X_i-\mu)^2}{\sigma^2}\sim\chi^2_{(n)} \qquad \frac{\sum(X_i-\bar{X})^2}{\sigma^2}\sim\chi^2_{(n-1)}}$$

$$E\left(\chi^2_{(n-1)}\right)=n-1 \qquad Var\left(\chi^2_{(n-1)}\right)=\sigma^2$$

ההסתברות שאנ"מ לשונות במדגם נורמאלי יסטה ממנה בסטייה יחסית שינה עולה על ε :

$$\hat{\sigma}^2=\frac{1}{n}\sum(X_i-\mu)^2 \qquad P\left(\left|\frac{\hat{\sigma}^2-\sigma^2}{\sigma^2}\right|\le\varepsilon\right)$$

$$=\boxed{P(W_n\le n(1+\varepsilon))-P(W_n\le n(1-\varepsilon))}$$

התפלגות t:

$$\boxed{T=\frac{Z}{\sqrt{W_k/k}}\sim t_{(k)}}$$

$$E(T)=0,\ Var(T)=k(k-2)$$

תכונות:
1. סימטרית סביב 0.
2. זנבות עבים משל הנורמאלית.

כאשר השונות לא ידועה: $\frac{\bar{X}-\mu}{S/\sqrt{n}}\sim t_{(n-1)}$

$$\boxed{t_{(n),0.1}=-t_{(n),0.9}}$$

התפלגות F:

$$\boxed{F(k,m)=\frac{W_k/k}{W_m/m}} \qquad E(F)=\frac{k}{m-2}$$

$$Var(F)=\frac{2m^2(k+m-2)}{k(m-2)^2(m-4)}$$

תכונות:
1. התפלגות לא סימטרית עם ד"ח במונה ובמכנה.
2. $F(1,k)\sim\left(t_{(k)}\right)^2$ .3 $\frac{1}{F(k,m)}\sim F(m,k)$

$$\boxed{F_{0.05}^{(k,m)}=\frac{1}{F_{0.95}^{(m,k)}}}$$

בראש הטבלה: a. בפנים: הסתברות-p.

$P(X\le a)=p$

סטטיסטיקה תיאורית ושיטות גרפיות:

שגיאת דגימה: שגיאה כתוצאה מכך שנאספו נתונים על מדגם שהינו חלק מכלל האוכלוסייה. גודלה תלוי בשיטת הדגימה ובגודל המדגם.

ממוצע קטום: ממוצע ללא הנמוך והגבוה.

שברון p באוכלוסייה: $P(X\le\xi_p)=p\ ,\xi_p$

הערך ש- p% מהאוכלוסייה מתחתיו. $F_X(\xi_p)=p$

חישוב השברון המדגמי:

$$r=(n+1)p-\lfloor(n+1)p\rfloor \qquad k=\lfloor(n+1)p\rfloor$$

$$Q(p)\to\boxed{(1-r)x_{(k)}+rx_{(k+1)}}$$

רבעון ראשון-Q1:

$$Q_1=\frac{3}{4}X_{\left(\left\lfloor\frac{n+1}{4}\right\rfloor\right)}+\frac{1}{4}X_{\left(\left\lfloor\frac{n+1}{4}\right\rfloor+1\right)}$$

חציון-M:

זוגי: $M=\frac{1}{2}X_{\left(\left\lfloor\frac{n+1}{2}\right\rfloor\right)}+\frac{1}{2}X_{\left(\left\lfloor\frac{n+1}{2}\right\rfloor+1\right)}$

אי-זוגי: $M=X_{\left(\left\lceil\frac{n}{2}\right\rceil\right)}$

תחום בין רבעונים:

$$IQR=Q_3-Q_1$$

רבעון שלישי-Q3:

$$Q_3=\frac{1}{4}X_{\left(\left\lfloor\frac{3(n+1)}{4}\right\rfloor\right)}+\frac{3}{4}X_{\left(\left\lfloor\frac{3(n+1)}{4}\right\rfloor+1\right)}$$

שונות המדגם: $S^2=\frac{1}{n-1}\sum(X_i-\bar{X})^2$

BOX-PLOT:

$$step=\frac{3}{2}(Q_3-Q_1)$$

$$\boxed{LF=Q_1-step}\ \boxed{UF=Q_3+step}$$

LW: התצפית המינימאלית שגדולה מ-LF.
UW: התצפית המקסימאלית שקטנה מ-UF.

מומנטים הם עקיבים וח"ה, וגם אמד שהוא פו' רציפה שלהם.

אמידה נקודתית:

סטטיסטי: פונקציה של התצפיות במדגם.
פרמטר: גודל שהוא מאפיין מסכם של האוכלוסייה.
אמד: סטטיסטי שאיתו אומדים את הפרמטר.
אומדן: הערך של האמד, המחושב מהמדגם.

שיטת המומנטים:

המומנט ה-k באוכלוסיה: $\mu_k=E(X^k)$

המומנט ה-k המדגמי: $M_k=\frac{1}{n}\sum X_i^k$.

השיטה: מבטאים את הפרמטר ע"י המומנטים של האוכלוסייה ואז מציבים במקום את המומנטים של המדגם

דוגמא: $f(x|\theta)=(\theta+1)x^{\theta},\ 0\le x\le 1$

$$\mu_1=E(X)=\int_0^1 x(\theta+1)x^{\theta}dx=\frac{\theta+1}{\theta+2}$$

$$\theta=\frac{1}{1-\mu_1}-2\to\boxed{\hat{\theta}=\frac{1}{1-M_1}-2=\frac{1}{1-\bar{X}}-2}$$

אמידת התוחלת (הפרמטר הוא- μ):

$$\boxed{\hat{\theta}=M_1=\bar{X}}\ \boxed{E(\hat{\theta})=\mu}\ \boxed{Var(\hat{\theta})=\frac{1}{n}Var(X_1)}$$

אמידת השונות (הפרמטר הוא σ^2):

$$\boxed{\hat{\theta}=S^2=\frac{1}{n}\sum(X_i-\bar{X})^2}\ \boxed{E\left(\frac{n}{n-1}\hat{\theta}\right)=\sigma^2}$$

שיטת הנראות המירבית:

בחירת הערך p עבורו ההסתברות של תוצאות המדגם היא הגדולה ביותר.
פונקציית נראות: פו' צפיפות משותפת-

$$L(\theta|x_1..x_n)=f_{\theta}(x_1..x_n)=\prod f_{\theta}(x_i)$$

מוצאים אמד שימקסם את $L(\theta)$ ע"י גזירה והשוואה ל-0.

תכונות:
1. אנ"מ אינו בהכרח יחיד.
2. לפעמים קל יותר למקסם את $l(\theta)=\ln L(\theta)$
3. אם $\hat{\theta}$ אנ"מ ל- θ אז לכל פו $\tau(\theta)$, אנ"מ ל- $\tau(\theta)$ יהיה: $\tau(\hat{\theta})$.

דוגמא:

$$L(\theta)=\prod(\theta+1)x_i^{\theta}=(\theta+1)^n\left(\prod x_i\right)^{\theta}$$

$$l(\theta)=\ln(L(\theta))=n\ln(\theta+1)+\theta\sum\ln(x_i)$$

$$\frac{\partial L(\theta)}{\partial\theta}=\frac{n}{\theta+1}+\sum\ln(x_i)=0\to\boxed{\hat{\theta}=-\frac{n}{\sum\ln x_i}-1}$$

כאשר x מוגבל ע"י הפרמטר, מכניסים אינדיקטור

דוגמא:

$$L(\theta)=\prod\lambda e^{-\lambda(x_i-\theta)}I\{x_i\ge\theta\}$$

$$=\lambda^n e^{-\lambda\sum(x_i-\theta)}I\{X_{(1)}\ge\theta\}$$

$$=e^{n\lambda\theta}I\{X_{(1)}\ge\theta\}\lambda^n e^{-\lambda\sum x_i}\to\boxed{\hat{\theta}=X_{(1)}}$$

חוסר הטייה: ממוצע האומדנים יהיה שווה בקירוב לערך הפרמטר אותו רוצים לאמוד. $\hat{\theta}$

אמד חסר הטייה אם: $\boxed{E(\hat{\theta})=\theta}$.

השגיאה הריבועית הממוצעת-תוחלת הסטייה:

$$MSE_{\hat{\theta}}(\theta)=E(\hat{\theta}-\theta)^2$$

$$=Var(\hat{\theta})+\left(E(\hat{\theta})-\theta\right)^2$$

נעדיף את האמד שבו ה-MSE הקטן ביותר.

בהתפלגות אחידה: האנ"מ עדיף ל $n\ge 3$ אבל הוא אמד מוטה. אח"ה: $\theta^*=\frac{n+1}{n}X_{(n)}$ אבל עדיין

לפי ה-MSE: $MSE_{\hat{\theta}}(\theta)\approx MSE_{\theta^*}(\theta)$

עקיבות: אומד שה-MSE שלו שואף ל-0.

$\boxed{MSE\xrightarrow[n\to\infty]{}0}$ $\boxed{\hat{\theta}\xrightarrow[n\to\infty]{}\theta}$ אם לכל $\varepsilon>0$ האמד

שואף לפרמטר בהסת' 1: $\boxed{\lim_{n\to\infty}P(|\hat{\theta}-\theta|<\varepsilon)=1}$

הערה: בד"כ כשה-MSE לא שואף ל-0, לא עקיב.

משפט: אם $\hat{\theta}$ עקיב אז פו' רציפה $\tau(\hat{\theta})$ עקיב

אם $\hat{\theta}$ אח"ה אז פו' ליניארית $\tau(\hat{\theta})$ אח"ה.

סטטיסטי מספיק:

פונקציה שבה כל המידע במדגם ביחס לפרמטר.

הגדרה: $S=S(X_1,,,X_n)$ סטטיסטי מספיק

אם"ם הצפיפות $f(X_1..X_n|S=s)$ או התפלגות

$P(X_1=x_1..X_n=x_n|S=s)$ לא תלויה ב θ , $\forall s$.

תכונות:
1. אם $\hat{\theta}$ אנ"מ ל θ אז $\hat{\theta}$ הוא פו' של ס"מ S .
2. ס"מ אינו יחיד, אם $S=g(S^*)$, גם: S^*
ס"מ וגם $f(S)$ ס"מ כאשר f חח"ע.

משפט הפירוק: סטטיסטי S מספיק ביחס ל θ

אם $f(x_1..x_n|\theta)=h(x_1..x_n)g(S(x_1..x_n)|\theta)$

כלומר: מפרקים את f ל- h שתלויה רק בתצפיות ו- g שתלויה ב- θ , כאשר S היא המידע מהתצפיות בתוך g .

משפט: אנ"מ תלוי במדגם מקרי דרך ס"מ.

דוגמא:

$$\underbrace{\left(\frac{1}{\sigma}\right)^n e^{-\frac{\sum x_i}{\sigma}}}_{h(\underline{x})}\cdot\underbrace{e^{\frac{\mu n}{\sigma}}I\{X_{(1)}\ge\mu\}}_{g(X_{(1)}|\mu)}\to\boxed{S=X_{(1)}}$$

משפט Rao-Blackwell: $\hat{\theta}$ אמד ו- S ס"מ:

1. $\theta^*=E(\hat{\theta}|S)$ הוא אמד. 2. $E(\theta^*)=E(\hat{\theta})$.
3. $MSE_{\theta^*}(\theta)\le MSE_{\hat{\theta}}(\theta)$ (שוויון - $\hat{\theta}$ כבר מבוסס על הס"מ).

דוגמא: $Ber(p),\exp(\lambda)$ אומד: X_1 . שיפורו:

$$E\left(X_1|\sum x_i\right)=P\left(X_1=1|\sum x_i=s\right)=\frac{\sum x_i}{n}=\bar{X}$$

$Pois(\lambda)$:

$$E(X_1|S)={}^{s}\!/_{n}\left({}^{n-1}\!/_{n}\right)^{s-1}$$

$$={}^{s}\!/_{n}\left(1-{}^{1}\!/_{n}\right)^{s-1}\sim Bin\left(s,{}^{1}\!/_{n}\right)$$

בדיקת השערות:

- בעיות שיש להחליט בהן על אחת מבין 2 השערות.

השערת האפס: ההשערה השמרנית, האמונה הרווחת.
האלטרנטיבה: ההשערה החדשנית, המהפכנית.

- החלטה שהשערה נכונה ע"י הוכחה שההשערה ההפוכה לא נכונה.
- מקבלים תוצאה קיצונית תחת השערת האפס משתי הסיבות הבאות:

1. מדגם לא מייצג 2. השערת האפס לא נכונה.

כלל החלטה: חלוקת הערכים האפשריים של תוצאות המדגם לשני תחומים:

1. R-אזור דחיית השערת האפס.
2. R^c- אזור אי-דחיית השערת האפס.

	אי-דחיית H_0	דחיית H_0
H_0 נכונה	V	טעות מסוג I
H_1 נכונה	טעות מסוג II	V

טעות מסוג I: דחיית השערת האפס כאשר היא נכונה.

$$\alpha=P(type\ I\ error)=P(rejecting\ H_0|H_0\ is\ true)$$

טעות מסוג II: קבלת השערת האפס כאשר אינה נכונה.

$$\beta=P(type\ II\ error)=P(not\ rejecting\ H_0|H_1\ is\ true)$$

הקטנת α תביא לעליה ב-β ולהפך.
השערה פשוטה: השערה הקובעת באופן יחיד את התפלגות האוכלוסייה שממנה נלקח המדגם.
השערה מורכבת: תוצאתה לא מאפשרת לקבוע את התפלגות האוכלוסייה.
פונקצית מבחן: לכל כלל הכרעה ישנה פונקציה שקובעת את ההסתברות שבה נדחה את השערת האפס עבור התוצאה x. $\alpha=E_{H_0}(\phi_D(X)),\ \beta=E_{H_1}(1-\phi(X))$

רמת מובהקות: ההסתברות המרבית המותרת לטעות סוג I.
רמת המובהקות שנקבעת תלויה במידת הנזק שייגרם כתוצאה מטעות מסוג I, אם הנזק גדול- נקבעת רמת מובהקות נמוכה.
הערה: קובעים α מראש ומחפשים עבורו β קטן.
עוצמה של מבחן: ההסתברות לדחיית השערת האפס כאשר האלטרנטיבה נכונה. $\boxed{\pi=1-\beta=P_{H_1}(reject\ H_0)}$

מבחן טוב: α קטן ו-π גדול.
דוגמא:

$X\sim Bin(3,p),\ p\in\{1/4,1/2\},\ H_0:p=1/2,H_1:p=1/4$

בר"מ 1/8, מהו מבחן בעל עוצמה מירבית?
X-מספר הפעמים שהוא נופל על "עץ".
$P_{H_0}(x=3)=1/8,\ P_{H_0}(x=0)=1/8$. איזה מהם עדיף?

$\pi=P_{H_1}(x=3)=1/64,\ \pi=P_{H_1}(x=0)=\boxed{24/64}$.

המבחן עם אזור דחייה-0 עדיף על זה עם אזור דחייה-3.

הלמה של ניימן-פירסן (מבחן יחס הנראות):

מבחן בעל עוצמה מרבית לבדיקת ההשערות הפשוטות:
הינו: $H_0:X\sim f_0\ ;\ H_1:X\sim f_1$
1- דחיית השערת האפס
0- אי-דחייה.

$$\boxed{\phi(x)=\begin{cases}1 & \lambda(x)>k_\alpha\\ \tau_\alpha & \lambda(x)=k_\alpha\\ 0 & \lambda(x)<k_\alpha\end{cases}}\qquad \boxed{\lambda(x)=\frac{f_1(x)}{f_0(x)}}\leftarrow \text{סטטיסטי}$$

λ גבוה: f_1 יותר מתאים. λ נמוך: f_0 יותר מתאים.
מ"מ בדיד: $\lambda(x)=k_\alpha$: נדחה את ה שערת האפס בסיכוי τ_α.
מ"מ רציף: ישנם רק ערכים 1 ו-0.
דוגמא: מדגם בגודל 10 מ- $\exp(\lambda)$. $H_0:\lambda=1,H_1:\lambda=2$

$$\lambda(\bar{x})=\frac{f_1(\bar{x})}{f_0(\bar{x})}=\frac{\prod\lambda_1e^{-\lambda_1x_i}}{\prod\lambda_0e^{-\lambda_0x_i}}=\left(\frac{\lambda_1}{\lambda_0}\right)^n e^{-(\lambda_1-\lambda_0)\sum x_i}$$

$\lambda=\left(\frac{2}{1}\right)^{10}e^{-\sum x_i}>k_\alpha$ היא פו' יורדת של $\sum x_i$ ולכן

המבחן שקול ל: דחה H_0 אם: $\sum x_i<k_\alpha^*=C$. מציאת C:

$$\alpha=P_{H_0}(reject\ H_0)=P(\sum x_i<C)=P(2\lambda_0\sum x_i<2\lambda_0C)$$
$$=P(\chi^2_{(2n)}<2\cdot1\cdot C)=P(\chi^2_{(20)}<2C)=0.05\rightarrow 2C=\chi^2_{(20),0.05}$$

!: אם λ פונקציה יורדת של הסטטיסטי-X: דחייה- X<C.
אם λ פונקציה עולה של הסטטיסטי-X: דחייה- X>C.
הערה חשובה: השערת האפס היא היחידה שקובעת את אזור הדחייה ולכן הוא יישאר זהה גם לאלטרנטיבה אחרת.

ממ"פ מהתפלגות נורמאלית ושונות ידועה:

מדגם מ- $N(\mu,\sigma^2)$. $H_0:\mu=\mu_0,\ \ H_1:\mu=\mu_1$.

$$\lambda(x)=\exp\left(\frac{n\bar{x}(\mu_1-\mu_0)}{\sigma^2}+\frac{n(\mu_0^2-\mu_1^2)}{2\sigma^2}\right)>k_\alpha$$

אם $\mu_0<\mu_1$ דחייה- $\boxed{\bar{X}>\frac{\sigma}{\sqrt{n}}z_{1-\alpha}+\mu_0}$.

אם $\mu_1<\mu_0$ דחייה- $\boxed{\bar{X}<-\frac{\sigma}{\sqrt{n}}z_{1-\alpha}+\mu_0}$.

עוצמת המבחן גדלה כאשר:

1. ככל שההפרש $\mu_1-\mu_0$ גדול יותר.
2. ככל ש- n גדול יותר.
3. ככל שסטיית התקן של האוכלוסייה-σ קטנה.
4. גודל המדגם המינימאלי הדרוש לעוצמה ור"מ נתונות:

$$\boxed{n=\frac{(z_\pi+z_{1-\alpha})^2\sigma^2}{(\mu_1-\mu_0)^2}}$$

מבחן בעל עוצמה מרבית במידה שווה:

האלטרנטיבה מורכבת מכמה חלקים. אז לכל ערך שרירותי f_1 מהאלטרנטיבה, המבחן של ניימן-פירסן (שנקבע רק לפי השערת האפס) הוא הטוב ביותר: בעל עוצמה מרבית לכל השערה פשוטה מתוך האלטרנטיבה.
Pvalue: בהינתן תוצאת מדגם מסוימות,

1. ההסתברות לקבל תוצאה קיצונית לפחות כמו התוצאה שהתקבלה בניסוי, בהנחה שהשערת האפס נכונה.
2. רמת המובהקות הקטנה ביותר שעבורה נדחה את השערת האפס.

אם $\alpha\geq Pvalue$	נדחה בר"מ α
אם $\alpha<Pvalue$	לא נדחה בר"מ α

הערה: ככל שנגדיל את-n, Pvalue יקטן כי הערך שהתקבל במדגם יהיה יותר "חזק" ואז הוא ייחשב יותר קיצוני תחת השערת האפס.
דוגמא: התפלגות פואסון- השערה מורכבת:

$n=10,X_i\sim Pois(\lambda),\ \ H_0:\lambda=4,\ H_1:\lambda<4$

קירוב לנורמאלי: $\sum X_i\sim Pois(10\lambda)$,

$E(\sum X_i)=Var(\sum X_i)=10\lambda$

$\frac{\sum X_i-10\lambda}{\sqrt{10\lambda}}\sim N(0,1)$.

המבחן: דחייה- $\boxed{Z=\frac{\sum X_i-10\lambda_0}{\sqrt{10\lambda_0}}<-z_{1-\alpha}}$

בדיקת השערה דו-צדדית:

$H_0:\mu=\mu_0,\ H_1:\mu\neq\mu_0\ ,\ N(\mu,\sigma^2)$

מבחנים שונים עבורם נדחה את H_0:

1. $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}>Z_{1-\alpha/2}\ or\ \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}<-Z_{1-\alpha/2}$
2. ר"ס: $\mu_0\notin\bar{X}\pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
3. Pvalue= $P_{H_0}(Z>X)+P_{H_0}(Z<-X)$

בהפרש פרופורציות לפי דף השערות

הערה: אזור דחייה בהשערה דו-צדדית יהיה מורכב משני חלקים זהים בגודלם בכל התפלגות, סימטרית או לא.
(ולכן כאן: Pvalue= $2P_{H_0}(Z<-X)$).

חישוב עוצמת המבחן: $\pi=1-\beta$

$$\beta=P\left(Z\leq z_{1-\alpha/2}+\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)-P\left(Z\leq -z_{1-\alpha/2}+\frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}\right)$$

בדיקת שתי השערות מורכבות:

$H_0:\theta\in W\ ,\ H_1:\theta\notin W$

ר"מ: $\alpha=\sup_{\theta\in W}P_{H_0}(rejecting\ H_0)$

דוגמא: $N(\mu,\sigma^2)$ $H_0:\mu\leq\mu_0\ ,\ H_1:\mu>\mu_0$.

$$\alpha=\sup_{\theta\in W}P(\bar{X}>C)=1-\Phi\left(\frac{C-\mu_0}{\sigma/\sqrt{n}}\right)$$

ההסתברות לטעות מסוג I היא הכי גדולה כאשר $\mu=\mu_0$ בדיוק.

מבחן יחס נראות מוכלל: $0\leq\Lambda(x)\leq1$

לא בעל עוצמה מקסימאלית

$$\text{סטטיסטי}\rightarrow\boxed{\Lambda(x)=\frac{L_{H_0}}{L_{H_0\cup H_1}}=\frac{\sup_{\theta\in W}f_\theta(\bar{x})}{\sup_{\theta\in\Theta}f_\theta(\bar{x})}}$$

המבחן: דחייה- $\boxed{\Lambda(x)\leq k_\alpha}$ ככל ש- Λ יותר קטן: עדות יותר חזקה לדחיית H_0.

הערות:

- במבחן יחס נראות מציבים אנ"מ במכנה.
- מבחן י"נ מוכלל המונה והמכנה מתחלפים (לעומת ניימן-פירסן) ולכן גם האי שוויון.

ממ"פ מהתפלגות ברנולי:

$H_0:p=p_0\ ,\ H_1:p=p_1\ \ ,\ Ber(p)$

$$\lambda(x)=\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^x\left(\frac{1-p_1}{1-p_0}\right)^n>k_\alpha$$

אם $p_0<p_1$ דחייה- $\sum x>C_\alpha$.
אם $p_1<p_0$ דחייה- $\sum x<C_\alpha$.

דוגמא: $\alpha=0.05,n=10,p_0=1/2,p_1=3/4$
מציאת אזור הדחייה:

$P_{H_0}(X=10)+P(X=9)<0.05$

$+P_{H_0}(X=8)>0.05$

לא נכניס את-8 לאזור הדחייה. $R=\{9,10\}$.

חישוב τ_α:

$P(X=10)+P(X=9)+\tau_\alpha P(X=8)=0.05$

ההתפלגות האסימפטוטית של י"נ מוכלל:

$\boxed{\Lambda^*=-2\log\Lambda\ \underset{n\to\infty}{\sim}\ \chi^2_{(r)}}$ $r=$ #ד"ח

r=ההפרש בין מספר הפרמטרים הידועים לאלה שצריך לאמוד תחת השערת האפס.
מתי משתמשים? כאשר ההתפלגות המדויקת של Λ אינה ידועה.

המבחן: $\boxed{\Lambda(x)\leq k_\alpha\Leftrightarrow\Lambda^*\geq C_\alpha=\chi^2_{(r),1-\alpha}}$

דוגמא: $x_i\sim Pois(\lambda)$, $H_0:\lambda=1,\ H_1:\lambda\neq1$.

$$\Lambda=\frac{\prod e^{-\lambda_0}\lambda_0^{x_i}/x_i!}{\prod e^{-\bar{X}}\bar{X}^{x_i}/x_i!}=e^{-n(\lambda_0-\bar{X})}\left(\frac{\lambda_0}{\bar{X}}\right)^{\sum x_i}<k_\alpha$$

$$\rightarrow -2\left(-n(\lambda_0-\bar{X})+\sum x_i(\ln\lambda_0-\ln\bar{X})\right)>C_\alpha=\chi^2_{(1),1-\alpha}$$

טיב התאמה:

למשתנה K קטגוריות בהסתברויות: $p_1,\cdots,p_k$,
$\sum p_j=1$, n תצפיות ב"ת מסווגות לפי הקטגוריות.
x_j - מספר תצפיות בקטגוריה ה-i. $\sum x_j=n$ אז:

$\boxed{\bar{x}\sim Multi(n,p_1,\cdots,p_k)}$ $\boxed{\hat{p}_j=\frac{x_j}{n}}$

בדיקת ההשערות: $H_0:\bar{p}=\bar{p}_0\ ,\ H_1:\bar{p}\neq\bar{p}_0$,
$\bar{p}_0=(p_{10},\cdots,p_{k0})$. תחת H_0: $E(x_j)=np_{j0}$.

קטגוריה	A_1	A_2	⋯	A_K	סה"כ
שכיחות במדגם	X_1	X_2	⋯	X_K	n
שכיחות צפויה	np_{10}	np_{20}	⋯	np_{K0}	n

הערות:

- סכום ההסתברויות הוא-1 ולכן: ההסת' לקטגוריה האחרונה הינה 1 מינוס סכום כל האחרות.
- אם $E_j=np_{j0}<5$ מאחדים קטגוריות.

$\boxed{\Lambda(x)=\prod\left(\frac{np_{j0}}{x_j}\right)^{x_j}}$ $\boxed{\Lambda^*=2\sum x_i\ln\left(\frac{x_j}{np_{j0}}\right)\sim\chi^2_{(k-1)}}$

דוגמא: האם הקובייה הוגנת?

תוצאה	1	2	3	4	5	6	
שכיחות במידגם	37	41	30	27	32	33	200

p_i-ההסתברות לתוצאה i, $1\leq i\leq6$. n=200.

$H_0:\forall i\ p_i=1/6$. $H_1:\exists i\ p_i\neq1/6$

המונה של Λ : $P_{H_0}(x_1=37,\cdots,x_6=33)$

$$=\frac{200!}{37!\cdots33!}\left(\frac{1}{6}\right)^{37}\cdots\left(\frac{1}{6}\right)^{33}$$

המכנה של Λ : יש למצוא אנ"מים: $\hat{p}_i=\frac{x_i}{200}$

$$\Lambda=\frac{(1/6)^{37}\cdots(1/6)^{33}}{(37/200)^{37}\cdots(33/200)^{33}}=\prod\left(\frac{200-1/6}{x_i}\right)^{x_i}$$

$-2\log\Lambda\sim\chi^2_{(5)}$ אמדנו 5-פרמטרים כי השישי הוא 1 פחות הסכום של האחרים.

המבחן של פירסון:

סטטיסטי מבחן לבדיקת טיב התאמה של המדגם למודל הסתברותי: $\boxed{X_P^2=\sum\frac{(X_j-np_{j0})^2}{np_{j0}}}$ עבור מדגם מספיק גדול, אם H_0 נכונה אז: $X_p^2\sim\chi^2_{(k-1)}$.

המבחן: דחייה- $X_p^2>\chi^2_{(k-1),1-\alpha}$

הערה: גם כאן מאחדים קטגוריות בהן $E_j<5$.
דוגמא: האם ההתפלגות גיאומטרית?

$H_0:X\sim Geo(1/2),\ H_1:else,\ \ p_{j0}=P_{H_0}(X=j)=\frac{1}{2^j}$

דוגמא: האם ההתפלגות נורמאלית?
אמידת התוחלת והשונות ע"י אנ"מ: מחליטים בהתאם להשערה, למשל: $\hat{\mu}=76,\ \hat{\sigma}^2=12.72^2$ עבור ציונים.

$$[a_j,a_{j+1}):\hat{p}_{0j}=\Phi\left(\frac{a_{j+1}-\hat{\mu}}{\hat{\sigma}}\right)-\Phi\left(\frac{a_j-\hat{\mu}}{\hat{\sigma}}\right)$$

הערה חשובה: היה צורך לאמוד פרמטרים גם תחת H_0 ולכן מאבדים ד"ח. עבור אמידת r, #ד"ח: k-1-r.
טרנספורמציה:

$$\boxed{Y=g(x),\ \ f_Y(y)=f_X(g^{-1}(Y))\cdot\left|\frac{\partial g^{-1}(Y)}{\partial Y}\right|}$$

$$\boxed{\begin{array}{l}X=\frac{1}{\theta}x^{\frac{1}{\theta}-1},\ \ Y=-\ln(x),\ Y\sim\exp\left(\frac{1}{\theta}\right)\\ g^{-1}(Y)=e^{-Y}=X\ ,\ \ \frac{\partial g^{-1}(Y)}{\partial Y}=-e^{-Y}\end{array}}$$

בדיקת אי-תלות:

בדיקת קשר בין שני משתנים קטגוריים.
A-קטגוריית השורה, B-קטגוריית העמודה.
הנתונים מסוכמים בטבלת שכיחות:

	B_1	B_2	⋯	B_j	⋯	B_J	סה"כ
A_1	X_{11}	X_{12}	⋯	X_{1j}	⋯	X_{1J}	$X_{1.}$
A_2	X_{21}	X_{22}	⋯	X_{2j}	⋯	X_{2J}	$X_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮
A_i	X_{i1}	X_{i2}	⋯	X_{ij}	⋯	X_{iJ}	$X_{i.}$
⋮	⋮	⋮		⋮		⋮	⋮
A_I	X_{I1}	X_{I2}	⋯	X_{Ij}	⋯	X_{IJ}	$X_{I.}$
סה"כ	$X_{.1}$	$X_{.2}$	⋯	$X_{.j}$	⋯	$X_{.J}$	n

$H_0:P(A_i\cap B_j)=P(A_i)P(B_j)\ \forall i,j,\ H_1:else$

דרך פיתרון:
מחשבים מהי השכיחות הצפויה בכל תא בהנחה שהשערת האפס נכונה. ובונים טבלת שכיחות.

$$\hat{P}(A_i)=\frac{X_{i.}}{n},\ \ \hat{P}(B_j)=\frac{X_{.j}}{n}$$

$$\boxed{E_{H_0}(X_{ij})=n\hat{P}(A_i\cap B_j)=\frac{X_{i.}X_{.j}}{n}}$$

סטטיסטי: $\boxed{X_P^2=\sum_i\sum_j\frac{(X_{ij}-X_{i.}X_{.j}/n)^2}{X_{i.}X_{.j}/n}}$

המבחן: דחייה- $X_P^2>\chi^2_{((r-1)\times(c-1)),1-\alpha}$

r/c= מספר קטגוריות שורה/עמודה

חישוב Pval: $Pvalue=P\left(\chi^2_{(r-1)(c-1)}>X_P^2\right)$

טווח: $(1-\alpha_1)<Pval<(1-\alpha_2)$,

α_1,α_2 - ההסת' שמופיעות בשורה העליונה.

דוגמא- טיב התאמה עם כופלי לגרנז':

$H_0:X\sim Multi(200,p_1,3p_1,p_3,0.5p_3,p_5,p_6)$

$L=\frac{200!}{x_1!\ldots}p_1^{x_1}(3p_1)^{x_2}p_3^{x_3}(0.5p_3)^{x_4}p_5^{x_5}p_6^{x_6}$
$=C+3^{x_2}0.5^{x_4}p_1^{(x_1+x_2)}p_3^{(x_3+x_4)}p_5^{x_5}p_6^{x_6}$

$\ln(L)=C+(x_1+x_2)\ln(p_1)+(x_3+x_4)\ln(p_3)+$
$x_5\ln(p_5)+x_6\ln(p_6)$
$+\lambda(1-(4p_1+1.5p_3+p_5+p_6))$

$\frac{\partial\ln(L)}{\partial p_1}=\frac{x_1+x_2}{p_1}-4\lambda\rightarrow\boxed{\hat{p}_1=\frac{x_1+x_2}{4\lambda}}$

$\boxed{\hat{p}_3=\frac{x_3+x_4}{1.5\lambda}}$ $\boxed{\hat{p}_5=\frac{x_5}{\lambda}}$ $\boxed{\hat{p}_6=\frac{x_6}{\lambda}}$ $\lambda=n=200$

חישוב סטטיסטי פירסון:

1	2	3	4	5	6	n
20	64	40	25	25	26	200
0.10	0.31	0.22	0.11	0.12	0.13	
21	63	43.4	21.6	25	26	200

$X_P^2=1/21+1/63+11.56/43.4+11.56/21.6$

אמדנו 3 פרמטרים (האחרון נקבע לפי האחרים) ולכן: (k-1-r)=6-1-3=2.

דוגמא- העדפת מוצר:

100 איש נשאלו בסקר מוצר מועדף מתוך-2.
H_0: אין העדפה. $p=1/2$. ר"מ: $\alpha=0.05$

$X_i\sim Bin(100,p)\sim N(100p,100pq)$

עוצמה: $X\overset{H_0}{\sim}N(50,5^2)\ \ \ X\overset{H_1}{\sim}N(60,24)$

$\pi=P_{p=0.6}(|Z|>z_{1-\alpha/2})=1-P(|Z|\leq1.96)=$
$1-P\left(-1.96<\frac{\sum x_i-50}{5}<1.96\right)=1-$
$P\left(\frac{-1.96\cdot5+50-60}{\sqrt{24}}<\frac{\sum x_i-60}{\sqrt{24}}<\frac{1.96\cdot5+50-60}{\sqrt{24}}\right)=$
$1-P(-4.041<z<-0.04)=$
$1-\Phi(-0.04)+\Phi(-4.041)=0.516$

חישוב Pvalue בהשערה דו"צ:

לוקחים את הסטטיסטי בערך מוחלט: $|T|$ (ערך חיובי). ובודקים את ההסת' ליפול באזורי הדחייה:

$$Pvalue=P_{H_0}(t_{(n)}>|T|)+P_{H_0}(t_{(n)}<-|T|)$$
$$=2\cdot P_{H_0}(t_{(n)}>|T|)=2\cdot(1-P(t_{(n)}<|T|))$$

חישוב Pvalue בהשערה חד-צדדית:

$H_1:p>p_0\ \ Pval=P(t_{(n)}>T)=\frac{1}{2}Pval(duz)$

$H_1:p<p_0$
$Pval=P(t_{(n)}<T)=1-\frac{1}{2}Pval(duz)$

רגרסיה ליניארית פשוטה:

בעזרת מודלים של רגרסיה ניתן לחקור את הקשר בין שני משתנים הקשורים באופן לא דטרמיניסטי זוהי הרחבה של המודל הדטרמיניסטי הליניארי- $Y=\beta_0+\beta_1X$ למודל ליניארי הסתברותי אשר מאפשר לנבא את הערך של Y על-סמך הערך של X.

הנחות המודל הליניארי

$\boxed{\varepsilon\sim N\left(0,\sigma^2\right)}$ $\boxed{Y=\beta_0+\beta_1X+\varepsilon}$

ככל ש-σ^2 גדולה יותר, התצפיות יותר מפוזרות ומתרחקות מהקו.

יודעים את X ומנחשים את Y בצורה הטובה ביותר (Y לא נקבע בוודאות כי ישנו "רעש" מבחינה הסתברותית).

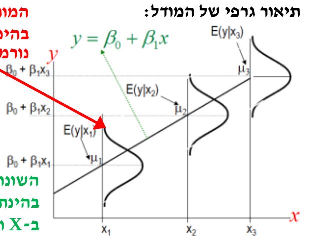
מתוך המצב רוצים לאמוד את β_0,β_1 כדי לנחש את Y בצורה טובה.

$$\boxed{E\left(Y|X=x\right)=E\left(\beta_0+\beta_1X+\varepsilon\right)=\beta_0+\beta_1X}$$

$$\boxed{Var\left(Y|X=x\right)=Var\left(\beta_0+\beta_1X+\varepsilon\right)=Var\left(\varepsilon\right)=\sigma^2}$$

Y- משתנה מוסבר. X- משתנה מסביר. לכל Y, ההתפלגות המותנית של Y בהינתן X היא נורמאלית

תיאור גרפי של המודל:



השונות של Y בהינתן X קבועה ב-X והיא עולה כפונקציה של X

משמעות σ^2 : ככל שהשונות יותר קטנה, נצפה שהתצפיות יהיו קרובות לישר הרגרסיה האמיתית.

אמידת מקדמי הרגרסיה: $\beta_0,\beta_1,\varepsilon$ $\quad \varepsilon_i\sim N\left(0,\sigma^2\right)$

מדגם מקרי: $\left(X_1,Y_1\right),\cdots,\left(X_n,Y_n\right)$, $Y_i=\beta_0+\beta_1X+\varepsilon_i$

אומדי ריבועים פחותים: הישר המביא למינימום את סכום ריבועי המרחקים האנכיים בין כל נקודה לישר.

$$\boxed{g\left(b_0,b_1\right)=\Sigma\left(Y_i-\left(b_0+b_1X_i\right)\right)^2}$$

אר"פ: $\hat\beta_0,\hat\beta_1$: ערכי b_0,b_1 המביאים למינימום את

מציאת אר"פ ע"י המשוואות הנורמאליות:

$$1)\ \frac{\partial g\left(b_0,b_1\right)}{\partial b_0}=-2\,\Sigma\left(Y_i-\left(b_0+b_1X_i\right)\right)\equiv 0$$

$$2)\ \frac{\partial g\left(b_0,b_1\right)}{\partial b_1}=-2\,\Sigma\,X_i\left(Y_i-\left(b_0+b_1X_i\right)\right)\equiv 0$$

$$1)\ \boxed{nb_0+b_1\Sigma X_i=\Sigma Y_i}\quad 2)\ \boxed{b_0\Sigma X_i+b_1\Sigma X_i^2=\Sigma X_iY_i}$$

פיתרון המשוואות הנורמאליות:

$$\boxed{\begin{aligned}b_0&=\hat\beta_0=\bar Y-\hat\beta_1\bar X\\ b_1&=\hat\beta_1=\frac{\Sigma\left(X_i-\bar X\right)\left(Y_i-\bar Y\right)}{\Sigma\left(X_i-\bar X\right)^2}=\frac{S_{XY}}{S_{XX}}\end{aligned}}$$

קו הרגרסיה עובר דרך נקודת הממוצעים $\left(\bar X,\bar Y\right)$

נוסחאות לאמידת קו הרגרסיה:

$$S_{XY}=\Sigma\left(X_i-\bar X\right)\left(Y_i-\bar Y\right)=\Sigma\left(X_i-\bar X\right)Y_i$$

$$=\Sigma X_i\left(Y_i-\bar Y\right)=\boxed{\Sigma X_iY_i-n\bar X\bar Y}$$

$$S_{XX}=\Sigma\left(X_i-\bar X\right)^2=\Sigma\left(X_i-\bar X\right)X_i=\boxed{\Sigma X_i^2-n\bar X^2}$$

$$S_{YY}=\Sigma\left(Y_i-\bar Y\right)^2=\Sigma\left(Y_i-\bar Y\right)Y_i=\boxed{\Sigma Y_i^2-n\bar Y^2}$$

$$S_X^2=S_{XX}/n-1\qquad S_Y^2=S_{YY}/n-1$$

האומד לישר הרגרסיה: $\boxed{\hat Y=\hat\beta_0+\hat\beta_1X}$

שאלות שניתן לענות עליהן:

- בהינתן X מסוים, מהו ה-Y המצופה?
פיתרון: ערך הנקודה Y על קו הרגרסיה.
- בהינתן הפרש של 1 ב-X מהו ההפרש המצופה ב-Y?
פיתרון: הניחוש להפרש יהיה השיפוע.
- תיאור החותך: הערך המצופה כאשר X=0 (אם X=0 לא בטווח התצפיות זוהי אקסטרפולציה)

אקסטרפולציה: ניבוי ערכי Y עבור ערכי X שאינם בטווח

תכונות האר"פ:

חסרי הטייה: $E\left(\hat\beta_1\right)=\beta_1$ $\quad E\left(\hat\beta_0\right)=\beta_0$

למציאת אר"פ אין שימוש בהנחת הנורמליות

עקיבים: $Var\left(\hat\beta_1\right)=\sigma^2\frac{1}{S_{XX}}$ $\quad Var\left(\hat\beta_0\right)=\sigma^2\frac{\Sigma X_i^2}{nS_{XX}}$

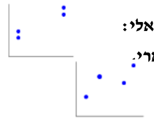
וגם אנ"מים.

ככל ששונות ה-X שבתצפיות יותר גדולה (המרחק ביניהם), יש יותר ביטחון בשיפוע הקו שמתקבל על-ידם.

דוגמא: יש לבחור 4 ערכי X שעבורם נמדוד את הקו.

פיתרון: נרצה X-ים שייתנו S_{XX} מקסימאלי כדי ש- $Var\left(\hat\beta_1\right)=\sigma^2/S_{XX}$ יהיה מינימאלי:

כאשר לא ידוע שהמודל הוא ליניארי. נבחר תצפיות יותר מפוזרות:



הגדרת שגיאה: $e_i=Y_i-\hat Y_i=Y_i-\hat\beta_0-\hat\beta_1X_i$

תכונות חשובות: $\boxed{\Sigma X_ie_i=0}$ $\boxed{\Sigma e_i=0}$

$e_i\neq\varepsilon_i$: e_i - מרחק התצפית מקו הרגרסיה.

ε_i - מרחק התצפית מהקו האמיתי.

הערך המנובא: לכל תצפית- $\hat Y_i=\hat\beta_0+\hat\beta_1X_i$.

השארית: לכל תצפית- $e_i=Y_i-\hat Y_i$

אמידת השונות במודל:

$$\left(Y|X=x\right)\sim N\left(\beta_0+\beta_1x,\sigma^2\right)$$

אמידת σ^2 על-סמך שונות השאריות:

$$\boxed{\begin{aligned}SSE&=\Sigma\left(e_i-\cancel{\bar e}_0\right)^2=\Sigma\left(Y_i-\hat Y_i\right)^2\\&=\Sigma\left(Y_i-\beta_0-\beta_1x_i\right)^2\quad \boxed{\hat\sigma^2=\frac{SSE}{n-2}}\end{aligned}}$$

n-2 בגלל אמידת החותך והשיפוע

טענה: $E\left(\hat\sigma^2\right)=\sigma^2$ (אמד חסר הטייה).

אמידת שונות האר"פ:

$$\boxed{Var\left(\hat\beta_0\right)=\hat\sigma^2\Sigma X_i^2/nS_{XX}}\quad\boxed{Var\left(\hat\beta_1\right)=\hat\sigma^2/S_{XX}}$$

אומדי נראות מירבית:

טענה: תחת הנחת הנורמליות

$\hat\beta_0,\hat\beta_1,\frac{n-2}{n}\hat\sigma^2$ הם אנ"מים ל- β_0,β_1,σ^2

פירוק סכומי הריבועים:

SST- סה"כ השונות: $\boxed{SST=SSR+SSE}$

SSR- השונות המוסברת, **SSE**- אינה מוסברת.

$$\boxed{\begin{aligned}SST&=\Sigma\left(Y_i-\bar Y\right)^2=S_{YY}\\SSR&=\Sigma\left(\hat Y_i-\bar Y\right)^2=\hat\beta_1^2S_{XX}=S_{XY}^2/S_{XX}\\SSE&=\Sigma e_i^2=\Sigma\left(Y_i-\hat Y_i\right)^2\end{aligned}}$$

מקדם המתאם הליניארי:

זהו מדד לתרומת המשתנה המסביר.

R^2: האומד לריבוע מקדם המתאם הליניארי.

זהו % השונות המוסברת של Y ע"י המודל.

$$0\le R^2\le 1\qquad \boxed{R^2=\frac{SSR}{SST}=1-\frac{SSE}{SST}}$$

מקדם המתאם: $-1\le\rho_{X,Y}=\frac{Cov\left(X,Y\right)}{\sigma_X,\sigma_Y}\le 1$

עבור X,Y ב"ת: $\rho_{X,Y}=0$ (ענן חסר צורה).

הערה: $\rho_{X,Y}=0$ לא גורר X,Y ב"ת!

עבור X,Y המקיימים: $Y=\beta_0+\beta_1X$

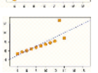
$$\rho_{X,Y}=\frac{\overset{\text{Cov(X,Y)}}{\beta_1\sigma_X^2}}{\underset{\sigma_Y}{|\beta_1|\sigma_X\sigma_Y}}=\begin{cases}1 & \beta_1>0\\-1 & \beta_1<0\end{cases}$$

אומד למקדם המתאם:

$$\boxed{R=\frac{\Sigma\left(X_i-\bar X\right)\left(Y_i-\bar Y\right)}{\sqrt{\Sigma\left(X_i-\bar X\right)^2}\sqrt{\Sigma\left(Y_i-\bar Y\right)^2}}=\frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}}$$

תכונות:

1. ערך R סימטרי עבור X ו-Y.
2. ערך R לא תלוי ביחידות המדידה של X,Y.
3. R≈0 מעיד על חוסר קשר ליניארי (ייתכן שישנו קשר שאינו ליניארי).
4. כאשר הקשר ליניארי, |R| מודד את חוזקו.
5. יכול להתקבל R גבוה גם כאשר הקשר לא ליניארי.
6. R רגיש לתצפיות חריגות (מודל הרגרסיה מניח שאין תצפיות חריגות).



הערה: לא ניתן להסתמך על R^2 גבוה לקביעת ליניאריות. ניתן להסתמך על נמוך לקביעת חוסר ליניאריות.

משוואת הרגרסיה:

$$\hat\beta_1=\frac{S_{XY}}{S_{XX}}=\frac{R\sqrt{S_{XX}S_{YY}}}{S_{XX}}=R\frac{S_Y}{S_X}$$

ביטוי קו הרגרסיה: $\boxed{\frac{Y-\bar Y}{S_Y}=R\frac{X-\bar X}{S_X}}$

משמעות: תקנון X,Y. אם X גבוה מהממוצע אז Y יהיה רק פי R גבוה מהממוצע.

R<1 : Y יתקרב לממוצע – "רגרסיה".

R=1: אין רגרסיה לממוצע.

R=0: רגרסיה מלאה, הניבוי הוא ממוצע הY.

הערה: אם X סטיית-תקן אחת מהממוצע (S_X), נצפה ש-Y יהיה R סטיות תקן מהממוצע.

השונות הבלתי מוסברת:

$$\boxed{\hat\sigma=\sqrt{\frac{\left(1-R^2\right)S_{YY}}{n-2}}=\sqrt{\frac{n-1}{n-2}}\sqrt{\left(1-R^2\right)}S_Y\approx\boxed{\sqrt{\left(1-R^2\right)}S_Y}}$$

הסטייה הטיפוסית מקו הרגרסיה

הסטייה מהקו של ערך Y בהינתן ערך X.

R<1: הסטייה מהקו קטנה מהסטייה הטיפוסית של Y.

|R|=1: הסטייה מהקו שווה לאפס.

(השונות במודל קטנה ככל ש-R^2 גדל)

R=0: הסטייה מהקו .

בדיקת השערות ורווחי סמך:

בדיקת השערות על השיפוע:

$H_0:\beta_1=\beta_1^0$	vs	$H_{1A}:\beta_1>\beta_1^0$
$H_0:\beta_1=\beta_1^0$	vs	$H_{1B}:\beta_1<\beta_1^0$
$H_0:\beta_1=\beta_1^0$	vs	$H_{1C}:\beta_1\neq\beta_1^0$

תוצאות חשובות לבניית סטטיסטי המבחן:

$$E\left(\hat\beta_1\right)=\beta_1,\ V\hat ar\left(\hat\beta_1\right)=\hat\sigma^2/S_{XX}$$

טענה: $\boxed{\left(n-2\right)\sigma^2/\sigma^2\sim\chi^2_{(n-2)}}$

טענה: $\hat\beta_1$ מתפלג נורמלית וב"ת ב- $\hat\sigma^2$.

סטטיסטי המבחן: $\boxed{T=\frac{\hat\beta_1-\beta_1^0}{\hat\sigma/\sqrt{S_{XX}}}\underset{H_0}{\sim}t_{(n-2)}}$

אם σ ידועה: $\boxed{\frac{\hat\beta_1-\beta_1^0}{\sigma/\sqrt{S_{XX}}}\underset{H_0}{\sim}N\left(0,1\right)}$

A: דוחים אם $T>t_{(n-2),1-\alpha}$

B: דוחים אם $T<-t_{(n-2),1-\alpha}$

C: דוחים אם $T>t_{(n-2),1-\alpha/2}$ או $T<-t_{(n-2),1-\alpha/2}$

$A:Pvalue=P(T\ge t)$

$B:Pvalue=P(T\le t)$

$C:Pvalue=P(|T|\ge|t|)$

ר"ס לשיפוע (ר"ס 1-α): $\boxed{\hat\beta_1\pm\frac{\hat\sigma}{\sqrt{S_{XX}}}t^{(n-2)}_{1-\alpha/2}}$

מקרה פרטי: $H_0:\beta_1=0,\ \ H_{1C}:\beta_1\neq0$

$$\boxed{T^2=\frac{\hat\beta_1^2S_{XX}}{\hat\sigma^2}=\frac{MSR}{MSE}\underset{H_0}{\sim}F_{(1,n-2)}}$$

זהו הסטטיסטי בעמודה F ב-ANOVA.

בדיקת השערות על החותך:

$H_0:\beta_0=\beta_0^0$	vs	$H_{1A}:\beta_0>\beta_0^0$
$H_0:\beta_0=\beta_0^0$	vs	$H_{1B}:\beta_0<\beta_0^0$
$H_0:\beta_0=\beta_0^0$	vs	$H_{1C}:\beta_0\neq\beta_0^0$

סטטיסטי המבחן:

$$\boxed{T=\frac{\hat\beta_0-\beta_0^0}{\hat\sigma\sqrt{\left(\Sigma X_i^2\right)/nS_{XX}}}\underset{H_0}{\sim}t_{(n-2)}}$$

אם σ ידועה:

$$\boxed{\frac{\hat\beta_0-\beta_0^0}{\sigma\sqrt{\left(\Sigma X_i^2\right)/nS_{XX}}}\underset{H_0}{\sim}N\left(0,1\right)}$$

A: דוחים אם $T>t_{(n-2),1-\alpha}$

B: דוחים אם $T<-t_{(n-2),1-\alpha}$

C: דוחים אם $T>t_{(n-2),1-\alpha/2}$ או $T<-t_{(n-2),1-\alpha/2}$

$A:Pvalue=P(T\ge t)$

$B:Pvalue=P(T\le t)$

$C:Pvalue=P(|T|\ge|t|)$

ר"ס לחותך: $\boxed{\hat\beta_0\pm\hat\sigma\sqrt{\frac{\Sigma X_i^2}{nS_{XX}}}t^{(n-2)}_{1-\alpha/2}}$

האמדים לחותך ולשיפוע מתפלגים נורמאלית:

$$\boxed{\hat\beta_0\sim N\left(\beta_0,\sigma^2\frac{\Sigma X_i^2}{nS_{XX}}\right)\quad\hat\beta_1\sim N\left(\beta_1,\frac{\sigma^2}{S_{XX}}\right)}$$

$$\boxed{\boxed{Cov\left(\hat\beta_0,\hat\beta_1\right)=-\sigma^2\frac{\bar X}{S_{XX}}}\ Cov\left(\hat\beta,\bar Y\right)=0\\ \forall i:\ Cov\left(\hat\alpha,e_i\right)=0,\ Cov\left(\hat\beta,e_i\right)=0}$$

ניתן להציג את האמד לחותך בצורה הבאה:

$$\hat\alpha=\Sigma c_iY_i,\quad c_i=\frac{1}{n}-\bar X\frac{X_i-\bar X}{S_{XX}}$$

פלט הרגרסיה:

Regression Statistics	
Multiple R	R
R Square	R^2
Adjusted R Square	
Standard Error	$\hat\sigma$
Observations	n

ANOVA Table

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F	Pvalue
רגרסיה	1	SSR	MSR=SSR	MSR/MSE	P(F>f)
שארית	n-2	SSE	MSE=SSE/(n-2)		
Total	n-1	SST			

F: סטטיסטי המבחן לבדיקת ההשערה: H_0: $\beta_1=0$.

Pvalue קטן: עדות חזקה לכך שיש קשר בין X ל-Y.

Variable	Parameter Estimate	Standard Error	T for H_0: $\beta_{0/1}=0$ H_1: $\beta_{0/1}\neq0$	Pvalue
חותך	$\hat\beta_0$	$\hat\sigma_{\hat\beta_0}=\hat S_Y=\hat\sigma\sqrt{\Sigma X_i^2/nS_{XX}}$	$\hat\beta_0-\beta_0^0/\hat\sigma_{\beta_0}$	עבור אותן השערות: ההסתברות להיות גדול או קטן מ-T
שיפוע	$\hat\beta_1$	$\hat\sigma_{\hat\beta_1}=\hat S_X=\hat\sigma/\sqrt{S_{XX}}$	$\hat\beta_1-\beta_1^0/\hat\sigma_{\beta_1}=\sqrt{F}$	

Pvalue בשורה של השיפוע שווה ל-Pvalue ב-ANOVA.

אמידת התוחלת של Y בהינתן X:

אמידת $E\left(Y|X=x_0\right)=\beta_0+\beta_1x_0\equiv\eta\left(x_0\right)$

ע"י: $\boxed{\hat\eta\left(x_0\right)=\hat\beta_0+\hat\beta_1x_0}$

התפלגות האמד:

-חסר הטייה $E\left(\hat\eta\left(x_0\right)\right)=\eta\left(x_0\right)$

$$Var\left(\hat\eta\left(x_0\right)\right)=\frac{\sigma^2}{n}+\left(x_0-\bar X\right)^2Var\left(\hat\beta_1\right)$$

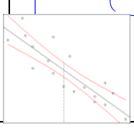
$$\cancel{+2Cov\left(\bar Y,\hat\beta_1\right)}_0=\sigma^2\left(\frac{1}{n}+\frac{\left(x_0-\bar X\right)^2}{S_{XX}}\right)$$

הערה חשובה: שונות מינימאלית- $x_0=\bar X$.

$$\boxed{\hat\eta\left(x_0\right)\sim N\left(\beta_0+\beta_1x_0,\sigma^2\left(\frac{1}{n}+\frac{\left(x_0-\bar X\right)^2}{S_{XX}}\right)\right)}$$

הנחות לבניית רווח תחזית:

- קשר ליניארי.
- שונות סטיות קבועה לכל X .
- אין תצפיות חריגות.
- התפלגות הסטיות נורמאלית.



ר"ס לתוחלת של Y בהינתן X:

$$\boxed{E\left(Y|X=x_0\right)\in\left[\hat\beta_0+\hat\beta_1x_0\pm\hat\sigma\sqrt{\frac{1}{n}+\frac{\left(x_0-\bar X\right)^2}{S_{XX}}}\cdot t^{(n-2)}_{1-\frac{\alpha}{2}}\right]}$$

הערה: ר"ס צר ביותר יתקבל עבור $x_0=\bar X$.

רווח תחזית:

בניית רווח עבור תצפית חדשה y_0 בהינתן x_0, שמרכזו על קו הרגרסיה והוא סימטרי סביבו.

הנחה: $\left(Y_0|X=x_0\right)\sim N\left(\beta_0+\beta_1x,\sigma^2\right)$

$$\boxed{\hat\beta_0+\hat\beta_1x_0\pm\hat\sigma\sqrt{1+\frac{1}{n}+\frac{\left(x_0-\bar X\right)^2}{S_{XX}}}\cdot t^{(n-2)}_{1-\frac{\alpha}{2}}}$$

$P\left(Y_0\in PI\right)=1-\alpha$ (PI- רווחת התחזית).

הערה: כאשר $n\to\infty$, ר"ס קטן והופך לנקודה, אבל אורך רווח תחזית שואף ל: $|2z_{1-\alpha/2}\sigma|$.

הערה: תמיד הר"ס מוכל ברווח תחזית. ככל שמתרחקים מ- $\bar X$ דיוק ר"ס פחות טוב, ואילו הרווח תחזית נשאר זהה (הקווים כמעט ישרים).