

生データの解析

Open cohort Studies

Open cohort study の典型的なものは下記のように表されます。

	exposed	unexposed	total
case	a	b	M_1
Person-time	N_1	N_0	T

Doll & Hill は喫煙の健康に対する効果を英国男性医師を対象に典型的 prospective study の形で調査しまとめています。

	喫煙	非喫煙	total
冠動脈疾患による死亡	630	101	731
Person-years	142,427	39,220	181,467

喫煙と冠動脈疾患による死亡の間に関連はあるのでしょうか？またあるとしたら、危険因子なのでしょうか、それとも抑制因子なのでしょうか？さらにその影響は大きいのでしょうか？小さいのでしょうか？

まずは仮説を証明します。

H_0 : 喫煙と冠動脈疾患死亡率の間に相関はない。

$$I_1 = I_0, I_1 - I_0 = 0, I_1 / I_0 = 1$$

H_A : 喫煙と冠動脈疾患死亡率の間に相関がある。

$$I_1 \neq I_0, I_1 - I_0 \neq 0, I_1 / I_0 \neq 1$$

I_1 : incidence rate in exposed pop, I_0 : incidence rate in non-exposed pop

基本は

$$Z^2 = \chi^2 = [X - E(X/H_0)]^2 / \text{var}(X/H_0)$$

$X = a = 630$, (実際観察された値)

$E(X/H_0) = M_1 N_1 / T = 731 \times 142,247 / 181,467 = 573.0$ (期待される値)

$$\text{var}(X/H_0) = M_1 N_1 / T (1 - N_1 / T) = M_1 N_1 N_0 / T^2 = 731 \times 142,247 \times 39,220 / 181,467^2 = 123.84$$

(binomial では $\text{var} = np(1-p)$ の形をとることを思いだしてください。)

$$Z^2 = \chi^2 = (630 - 573)^2 / 123.84 = 26.23$$

$$\Pr(\chi^2 > 26.23) = 0.000001$$

H_0 は棄却され、喫煙と冠動脈疾患死亡率の間に相関があると結論できます。もちろん confounder や bias が存在せず研究が valid であると仮定してのことです。

Rate difference (RD)

$$RD = a/N_1 - b/N_0 = 630/142,247PY - 101/39,220PY = 4.43/1,000PY - 2.58/1,000PY = 1.85/1,000PY$$

$$E(X/H_0) = E(RD/H_0) = 0$$

何故なら仮説 H0 において RD は 0 であると仮定しているからです。

$$\text{Var}(X/H_0) = \text{var}(RD/H_0) = I \times (1/N_1 + 1/N_0), \quad I = M_1/T \\ = 731/181,467 (1/142,247 + 1/39,220) = 0.00000013103$$

$$Z^2 = \chi^2 = (0.00185 \cdot 0)^2 / 0.00000013103 = 26.23$$

$$\Pr(\chi^2 > 26.23) = 0.000001$$

前と同じ結果を得ることができました。

次に 95%CI を算出します。

$$X \pm 1.96 \sqrt{\text{var}(X)}$$

$$X = RD = 0.0185 \cdot PY$$

$$\text{Var}(X) = \text{var}(RD) = I_1/N_1 + I_0/N_0 = 1/N_1^2 + b/N_0^2 = 630/142,247^2 + 101/39,220^2 \\ = 9.680/10^8/PY^2$$

$$1.85 \times 1,000PY \pm 1.96 \sqrt{9.680/10^8/PY^2} = (1.24/1000PY, 2.46/1000PY)$$

$$\text{Var}(RD) = \text{var}(X/H_0) = I \times (1/N_1 + 1/N_0)$$

何故なら H₀ では RD は 0 であると仮定しているからです。

Incidence Rate Ratio (IRR)

$$RR = (630/142,247PY) / (101/39,220PY) = 1.72$$

喫煙による 72% 冠動脈疾患による死亡が増加するといえます。

同様に 95%CI を算出します。

$$X \pm 1.96 \sqrt{\text{var}(X)}$$

$$X = \ln(RR) = \ln 1.72 = 0.5422$$

$$\text{Var}(X) = \text{var}[\ln(RR)] = 1/a + 1/b = 1/630 + 1/101 = 0.01149$$

$$0.5422 \pm 1.96 \sqrt{0.01149} = (0.3321, 0.7523)$$

$$95\%CI = e^{X \pm 1.96 \sqrt{\text{var}(X)}} = e^{(0.3321, 0.7522)} = (1.39, 2.21)$$

RDの場合と違って、RRではlnを使用して計算し、また最後にeを使ってもどしてきます。何故でしょうか？

RRは1が中央値です。範囲は0から無限大です。要するに負の関係がある場合数値は0 - 1をとり正の関係の場合には数値は1 - をとります。もしlnを使用すれば、1は0となり0は- となるので都合が良いのです。

以下は簡単な高校の数学の復習です。

$$\text{Exp}(X) = e^X$$

$$\text{Exp}(0) = 1$$

$$\text{Exp}(1) = 2.718281828$$

$$\text{Ln}[\text{exp}(x)] = X$$

$$\text{Ln}[1] = 0$$

$$\text{Ln}(2.718281828) = 1$$

$$\text{Ln}(2.718281828^2) = 2$$

$$\text{Ln } 2.718281828 = \text{Ln}[e^{0.5}] = 0.5$$

$$\text{Ln}[0] = -$$

$$\text{exp}[\text{Ln}(X)] = X$$

$$\text{exp}(A)\text{exp}(B) = \text{exp}(A+B)$$

$$\text{exp}(A)/\text{exp}(B) = \text{exp}(A-B)$$

$$\text{Ln}(AB) = \text{Ln}(A) + \text{Ln}(B)$$

$$\text{Ln}(A/B) = \text{Ln}(A) - \text{Ln}(B)$$

II. Case Control Study

Case control study は下のパターンになります。

	Exposed	Non-exposed	total
Disease	a	b	M_1
No-disease	c	d	M_0
total	N_1	N_0	T

下の表は経口避妊薬と心筋梗塞の関連をみた case control study です。

	経口避妊薬使用	使用せず	合計
心筋梗塞	23	112	135
なし	130	1306	1436
合計	153	1418	1571

Shapiro et al., Lancer 1979

経口避妊薬使用と心筋梗塞発生の間に関連はあるのでしょうか？またあるとしたら、危険因子なのでしょうか、それとも抑制因子なのでしょうか？さらにその影響は大きいのでしょうか？小さいのでしょうか？

まずは仮説を証明します。

H_0 : 経口避妊薬使用と心筋梗塞発生の中に相関はない。

OR = 1, RR = 1

H_A : 経口避妊薬使用と心筋梗塞発生の中に相関がある。

OR ≠ 1, RR ≠ 1

$$Z^2 = \chi^2 = [X - E(X/H_0)]^2 / \text{var}(X/H_0)$$

$X = a = 23$, (実際観察された値)

$E(X/H_0) = M_1 N_1 / T = 135 \times 153 / 1571 = 13.15$ (期待される値)

$\text{var}(X/H_0) = M_1 M_0 N_1 N_0 / T^2 (T-1) = 135 \times 1436 \times 153 \times 1418 / 1571^2 \times 1570 = 10.85$

$$Z^2 = \chi^2 = (23 - 13.15)^2 / 10.85 = 8.94$$

$$\Pr(\chi^2 > 8.94) = 0.003$$

H_0 は棄却され、経口避妊薬使用と心筋梗塞発生の中に相関があると結論できます。もちろん confounder や bias が存在せず研究が valid であると仮定してのことです。

$$\text{OR} = ad/bc = 23 \times 1306 / 112 \times 130 = 2.06$$

経口避妊薬使用により心筋梗塞の発生がおよそ2倍になることを示しています。

同様に 95% CI を算出します。

$$X \pm 1.96 \sqrt{\text{var}(X)}$$

$$X = \ln(\text{OR}) = \ln 2.06 = 0.7242$$

$$\text{Var}(X) = \text{var}[\ln(\text{OR})] = 1/a + 1/b + 1/c + 1/d = 1/23 + 1/112 + 1/130 + 1/1306 = 0.0609$$

$$0.7242 \pm 1.96 \sqrt{0.0609} = (0.241, 1.21)$$

$$95\% \text{CI} = e^{X \pm 1.96 \sqrt{\text{var}(X)}} = e^{(0.241, 1.21)} = (1.27, 3.35)$$

この結果は経口避妊薬の使用により95%の信頼性をもって1.3倍から3.4倍に増加すると言えます。

II. Closed Cohort Study

Closed cohort study は下のパターンになります。

	Exposed	Non-exposed	total
Disease	a	b	M_1
No-disease	c	d	M_0
total	N_1	N_0	T

下の表は経口避妊薬と心筋梗塞の関連をみた case control study です。

	カテコラミン高濃度	カテコラミン低濃度	合計
冠動脈疾患	237	44	71
なし	95	443	538
合計	122	487	609

Cassel et al. Arch Intern Med 1971

血中カテコラミン濃度の高低と冠動脈疾患発生の間に関連はあるのでしょうか？またあるとしたら、危険因子なのでしょうか、それとも抑制因子なのでしょうか？さらにその影響は大きいのでしょうか？小さいのでしょうか？

まずは仮説を証明します。

H_0 : 血中カテコラミン濃度の高低と冠動脈疾患発生の間に関連はない。

$C_1 = C_0, C_1 - C_0 = 0, C_1 / C_0 = 1$

H_A : 血中カテコラミン濃度の高低と冠動脈疾患発生の間に関連がある。

$C_1 \neq C_0, C_1 - C_0 \neq 0, C_1 / C_0 \neq 1$

$$Z^2 = \chi^2 = [X \cdot E(X/H_0)]^2 / \text{var}(X/H_0)$$

$$X = RD = C_1 - C_0 = a/N_1 - b/N_0 = 27/122 - 44/487 = 0.131$$

$$E(X/H_0) = E(RD/H_0) = 0$$

$$\text{Var}(X/H_0) = \text{var}(RD/H_0) = C(1-C)/N_1 + C(1-C)/N_0 = C(1-C)(1/N_0 + 1/N_1) \\ = 71/609(1 - 71/609)(1/122 + 1/487) = 0.00106$$

$$Z^2 = \chi^2 = (0.131 \cdot 0)^2 / 0.00106 = 16.25$$

$$\Pr(\chi^2 > 16.25) = 0.00006$$

H_0 は棄却され、血中カテコラミン濃度の高低と冠動脈疾患発生の間に関連があると結論できます。もちろんconfounder やbias が存在せず研究がvalidであると仮定してのことです。

同様に 95%CI を算出します。

$$X \pm 1.96 \sqrt{\text{var}(X)}$$

$$X = RD = 0.131$$

$$\text{Var}(X) = \text{var}(RD) = C_1(1 - C_1)/N_1 + C_0(1 - C_0)/N_0 = ac/N_1^3 + bd/N_0^3 \\ = 27 \times 95/122^3 + 44 \times 443/487^3 = 0.001582$$

$$0.131 \pm 1.96 \sqrt{0.001582} = (0.053, 0.209)$$

95% の信頼性をもって RD は 5.3% から 20.9% の間にあるといえます。

Cumulative Risk Ratio

$$RR = (27/122)/(44/487) = 2.45$$

カテコラミン濃度が高いと 2.5 倍冠動脈疾患になる率が上がることを示しています。

95%CI

$$X \pm 1.96 \sqrt{\text{var}(X)}$$

$$X = \ln(RR) = \ln 2.45 = 0.8959$$

$$\begin{aligned} \text{Var}(X) &= \text{var}[\ln(RR)] = (1 - C_1)/N_1C_1 + (1 - C_0)/N_0C_0 = c/aN_1 + d/bN_0 \\ &= 95/27 \times 122 + 443 / 44 \times 487 = 0.0495 \end{aligned}$$

$$0.8959 \pm 1.96 \sqrt{0.0495} = (0.4598, 1.332)$$

$$95\%CI \quad e^{X \pm 1.96 \sqrt{\text{var}(X)}} = e^{(0.4598, 1.332)} = (1.58, 3.79)$$

この結果はカテコラミン濃度の高低により 95%の信頼性をもって 1.6 倍から 3.8 倍に冠動脈疾患増加すると言えます。もちろん confounding, bias が存在しないと仮定してのことです。