疫学調査における統計学的推論の原理

Hypothesis testing　仮説テスト

ある疫学者が小児肥満は日本全体と比較して田舎に多いのではないかと仮説をたてました。この仮説を証明するため、彼女はある地方の３つの村に行き学校に登録された 8—12 歳の小児 355 人の body mass index を調査しました。ある cut-off 値を用いたところ 37%が肥満に属しました。一方日本全体でのこの年齢での肥満の割合は 30%とします。仮説をたて、統計学的に推論してみてください。

H_0：田舎の子供の肥満率(p_1)は日本全体(p_0)とかわらない。

$p_1 = p_0$

$X = 0.37 = p_1$

$E(x/H_0) = 0.30$

$Var\ (X/H_0) = p_0(1 - p_0) / n = 0.30\ (1 - 0.30) / 355 = 0.000592$

$\therefore Z^2 = [X - E(X/H_0)]^2 / var\ (X/H_0) = (0.37 - 0.3)^2 / 0.000592 = 8.28$

degree of freedom = 1 (2 x 2 table の場合 [2-1]x{2-1}=1 なので)

$Pr[\chi^2 > 8.28] = 0.004$

よってH_0は棄却され、‘日本全体と田舎では小児肥満の比率が異なる’と結論できます。

これは従来の生物統計の基本的手法です(生物統計学の基礎知識の項を参照)。しかし、この結論からは田舎では小児肥満が少ないともとれますし、統計学的にどれくらい違うのかがはっきりしません。χ^2テストは常にtwo-tailed なので多いか少ないかの推論ができないのです。

Confidence interval : CI（信頼区間）

信頼区間は疫学者に好んで用いられます。何故なら上記のような制限がないからです。

$95\%CI = x \pm 1.96\sqrt{var(X)}$

これは 95%の確率をもって本当の値がこの範囲に収まるということです。この場合、本当の値はわからず推論にすぎません。その推論は信頼区間が狭ければ狭いほど強力で

あり、広いと弱くなります。もし信頼区間がnull hypothesisの値Xを含まなければ、有意です。例えばratio の場合はH_0では１、difference の場合０になるはずです。ですからratio で９５％CIが1.5―2.0であれば、これだけで有意であるといえますが、(0.3,1.5)は1をまたいでいるので有意でありません。またdifferenceの場合(0.2, 0.6)は有意ですが-0.4, 0.8は0をまたいでいるので有意ではありません。

先ほどの小児肥満の例に戻って95％CIを出してみましょう。

$X \pm 1.96\sqrt{var(X)} = 0.37 \pm 1.96\sqrt{0.37(1-0.37)/355} = (0.32, 0.42)$

日本全体の0.30を含んでいませんので有意といえます。ここで注意していただきたいのは、p-value の際はH_0のもとに行なうわけですからvariance $(X/H_0) = 0.30(1-0.30)/355$ となります。しかし 95％CIの場合、H_0を仮定していないのでvar(X)を用います。

もう一度H_0に戻ります。

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

においてχ^2を大きくするには測定集団の平均xが変化して比較集団との隔たりが大きくなるか、$var(X/H_0)$が小さくなるかです。前者は生物学的特徴にとって規定されるため変化しません。$Var(X/H_0)$はRRとRDで異なります。

RR

$Var[\ln(RR)] = (1-C_1)/N_1C_1 + (1-C_0)/N_0C_0$

N_1=exposed された人数

N_0= exposed されなかった人数

Null hypothesis: $C_1 = C_0 = C$

$Var[\ln(RR/H_0)] = 1-C/C\ (1/N_1 + 1/N_0)$

RD

$Var[RD] = I_1/N_1 + I_0/N_0$

Null hypothesis: $I_1 = I_0 = I$

$Var[RD/H_0] = I\ (1/N_1 + 1/N_0)$

アメリカ人が20年間で大腸癌になる危険性は1%です。もしある物質Qに暴露された場合、20年間の危険性は1.1%になったとします。これは有意ととれますか。

統計学的には母集団の数によって有意でありえるのです。

N_1	N_0	N_1／{N_1＋N_0}	Var(RD/H_0)	χ^2	p-value
100	100	0.50	0.000198	0.005	0.94
1000	1000	0.50	0.0000198	0.050	0.82
1000	9000	0.10	0.000011	0.091	0.76
5000	5000	0.50	0.00000396	0.252	0.62
10,000	10,000	0.50	0.00000198	0.505	0.48
50,000	50,000	0.50	0.000000396	2.525	0.11
100,000	100,000	0.50	0.000000198	5.05	0.02

上の表をみて解る通り物体Qに暴露された人10万人と暴露されなかった人10万人を集めた結果であれば、かりに0.1%の増加であっても有意です。逆に数が少なければ、50%増えても有意でないこともありえるのです。またN_1とN_0の合計だけでなくN_1／{N_1＋N_0}にも影響されます。N_1＝N_0の方がχ^2は高くなります。何故ならRDの場合

$Var[RD] = I_1/N_1 + I_0/N_0$だからです。

χ^2＞3.84 の際有意であると覚えておくと便利です。

疫学研究におけるrandom variability の源

Randomized clinical trial

randomized clinical trial においては参加者を無作為に２つのグループに割り付けるため、理論上既知、未知の confounder は打ち消されるはずです。ですから、２つのグループで多少分布が異なるのは random imbalance ということになります。

Observational study

この研究に関しては exposure は均等に割り付けられません。よって両者の違いは単純な chance ではかたづけられないのです。

Sampling

物質 Q に対する作業員の暴露状況は工場のどの部署で働いたかによるとします。その配置が全く適当に割り振られていたとすればこれも無作為ということになります(**randomization by nature**)。これは研究者の経験によって発生したものでなく、所謂コインを投げて表か裏かの問題と同じです(**designed by nature**)。

<u>統計学的有意性は研究が有効に行なわれたという前提のもとに成り立つ</u>

bias, confounder が存在しない状態(internal validity)において統計学的に検討し有効なのです。もしも internally valid でないのにいくら統計学的に推論しても意味はありません。例えば問題を写し間違えて解答しても、いくらその間違った問題に対する解答が正しくても答えは間違っているのです。

もしも上の例で、男性作業員の方が女性作業員より物質 Q に暴露される機会が多いとします。一般的に男性の方が女性より冠動脈疾患に罹患しやすいことも判っています。よって性は confounder です。しかし男性にのみ的を絞れば、物質 Q への暴露はランダムであったことが予想されます。