SIMPON'S PARADOX

The value of a measure of association btw 2 var, may be identical w/in the levels of a third variable but can take an entirely ≠ value when this 3rd variable is disgregated and the associated is computed from pooled data

POLYTOMUS EXPOUSURE and OUTCOME guarda slide

- n° of parameters in saturated models $(I J) - 1$
- n° of parameters IF we assume indepemdemce $(I-1)+(J-1)$

we test the hypothesis of indepemdemce w/ X^2 test or chi squared

2 approaches: – Test indep. : btw expousure and outcome by comparing observ vs exp freq
↳ categorical
↳ RIDIT SCORES = compares 1 or + gr to a stramd. gr

RIDIT = ① Frequency A — RIVEDERE
② $\frac{1}{2}$ Frequency
③ cumulative frequency starting from zero
④ 2 + 3
⑤ 4/N_A

$\bar{R} = \frac{\text{sum of prod}}{N_B}$ { > 0.5 more likel; < 0.5 less likely

$\frac{\bar{R}}{1-\bar{R}}$ = ODDS

GENERALIZED LINEAR MODELS { linear, Bimomial, Poissom

① LOGISTIC MODEL

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

β_0 = log ODDS of success
$\exp(\beta_0)$ = ODDS of success

β_1 = slope, is the change in the log odds of success for + unit increase in X → log ODDS RATIO
$\exp \beta_1$ = ODDS RATIO

② LINEAR MODEL

$$\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

β_0 = average for baseline, or expected m

β_1 = ≠ btw 2 averages — change in mean
ratio btw 2 expected m w/ + unit increase

MEASURING PREDICTIVE POWER

AUC = area under the ROC curve that describes (y) sensitivity and 1 - specificity (x)
Also the proportion of concordant pair

$$\frac{C}{C+T+D}$$

$$\text{CStat } \frac{C + 0.5T}{C+T+D}$$

$$\text{gimi coeff } d = \frac{C-D}{C+T+D}$$

C: $y_1 = 1 \quad y_2 = 0 \quad \pi_1 > \pi_2$
D: $y_1 = 1 \quad y_2 = 0 \quad \pi_1 < \pi_2$
T: $y_1 = 1 \quad y_2 = 0 \quad \pi_1 = \pi_2$

TEST OF AGREEMENT (for matched data)

① INTER OBSERVERS AGREEMENT = in the diagonal we observe perfect agreement, outside = disagreement

sum of pr in the diagonal = 1

COHEN'S KAPPA = $\frac{\text{sum of the pr in the diag} - \text{sum pr under indep}}{1 - \text{sum pr independence}}$

$H_0: K = 0$ pr in the diag = pr ind
$K = 1$ num and den are the same, sum = 1

TEST OF HOMOGENEITY

① BD TEST : test of collapsibility, if all stratum specific have same OR or RR

$H_0: \psi_1 = \psi_2 = \bar{\psi}_K$ if so, they can be reported as weighted averages
if we reject BD test, the stratum specific estimates must be reported separately

② CMH TEST : used to test specific hypothesis

$H_0: \psi_1 = \psi_2 = \ldots = \psi_K = 1$

③ MC NEMAR TEST : for matched data, we test marginal homogeneity

$H_0: \pi_{1+} = \pi_{+1}$

(glm) $\rightarrow T = \frac{m_{21} - m_{12}}{\sqrt{m_{21} + m_{12}}}$ (clogit) $\rightarrow T = \frac{m_{12}}{m_{21}}$

if T > pval we reject H_0 and accept $H_1: \pi_{1+} \neq \pi_{+1}$

FACTORIAL NUMBERS

$$Pr(y) = \frac{m!}{y!(m-y)!} \cdot p^y (1-p)^{m-y}$$

STATISTICS FOR HEALTH

- Linear relationship — COVARIANCE $cov(x,y) = E\{(x-E(x))(y-E(y))\}$
  - if = 0 no linear rel
  - CORRELATION COEFFICIENT: −1 (neg lin rel ↑↓ ↓↑) · 0 (no corr) · +1 (pos lin rel ↑↑ ↓↓)

- ESTIMATION — SAMPLE VARIANCE $S^2 = \frac{\Sigma(y_i-\bar{y})^2}{m}$ correct $S^2 \Sigma \frac{(y_i-\bar{y})^2}{m-1}$
  - STANDARD ERROR SE: $\sqrt{\frac{\sigma^2}{m}} = \frac{\sigma}{\sqrt{m}}$
  - MAXIMUM LIKELY ESTIMATOR $L:(p) = [p^{y_1}(1-p)^{1-y_1}] \cdot \ldots \cdot [p^{y_{10}}(1-p)^{1-y_{10}}]$
  - POINT ESTIMATION — SAMPLE MEAN: $\bar{\mu} = \frac{\Sigma \mu_i}{m}$
    - SAMPLE VARIANCE: σ^2 or S^2 $\frac{\Sigma(y_i-\bar{y})^2}{m-1}$
  - INTERVAL ESTIMATION $CI = \beta \pm SE \cdot z_{1-\frac{\alpha}{2}}$

TEST ON THE MEANS

We have 1 population w/ mean μ, $H_0 = \mu = \mu_0$ and $H_1: \mu \neq \mu_0$
H_1 is a two sided hypothesis bc μ can be both > or < of μ_0

If we don't know σ^2 we can use the estimate of $S^2 = \frac{\Sigma(y_i-\bar{y})^2}{m-1}$ (df)

① TWO SIDED TEST ON THE MEAN: $\left|\frac{\bar{y}-\mu_0}{S/\sqrt{m}}\right| \geq t_{1-\frac{\alpha}{2},(m-1)}$ → critical value

if so we REJECT H_0

② ONE SIDED TEST ON THE MEAN: $\frac{\bar{y}-\mu_0}{S/\sqrt{m}} > t_{1-\alpha,(m-1)}$

If we have 2 pop or 2 independent samples (y_1 w/ σ_1^2, y_2 w/ σ_2^2)
We want to test the equality of the 2 means
$H_0: \mu_1 = \mu_2$ $H_1: \mu_1 \neq \mu_2$

We assume that $\sigma_1^2 = \sigma_2^2$ →

③ TEST OF EQUALITY IN 2 MEANS TWO SIDED: $\left|\frac{\bar{y}_1-\bar{y}_2}{Sp\cdot\sqrt{\frac{1}{m_1}+\frac{1}{m_2}}}\right| \geq t_{1-\frac{\alpha}{2},(m_1+m_2-2)}$

$Sp = \frac{(m_1-1)S_1^2+(m_2-1)S_2^2}{m_1+m_2-2}$ → pooled var

- if we assume $\sigma_1^2 \neq \sigma_2^2$ $\left|\frac{\bar{y}_1-\bar{y}_2}{\bar{S}}\right| \geq t_{1-\frac{\alpha}{2}\cdot(m)}$

$\bar{S} = \frac{S_1^2}{m_1} + \frac{S_2^2}{m_2}$

STUDY DESIGN

① COHORT STUDIES : Healthy population divided in E^+ and E^-
Calculate the OCCURENCE of disease to see if its the same btw E^+ and E^-
cons = time consuming, exp, no good for rare dis
Longitudinal

② CASE-CONTROL STUDIES : unhealthy (cases) vs healthy (controls) ppl
they are already ill
cons: can not estimate prevalence or independence
pro: good for rare dis, multiple exposure
RETROSPECTIVE

③ CROSS-SECTIONAL STUDIES: target population, strict criteria
pop divided in: D^+ D^- E^+ E^-
cons: only prevalence, no good for rare dis
pro: not expensive

RISK

- $RR = \frac{a/m_1}{c/m_0} = \frac{\lambda_1}{\lambda_0}$
- EXCESS RISK = $\lambda_1 - \lambda_0$
- Attributable risk = $\frac{\lambda_1 - \lambda_0}{\lambda_1}$
- Crude risk : π : $\frac{m^\circ \text{ cases}}{m \text{ pop}}$
- incidence rate : $\lambda = \frac{m \text{ cases}}{p \cdot \text{years}}$
- RISK = R : $1 - \exp(\lambda t)$
- ODDS RATIO: OR: $\frac{ad}{bc}$ $\quad OR = RR\left(\frac{1-\lambda_1}{1-\lambda_0}\right)$

$$\text{SENSITIVITY} = \frac{TP}{TP+FN}$$

$$\text{SPECIFICITY} = \frac{TN}{TN+FP}$$

$$PPV = \frac{TP}{TP+FP}$$

$$NPV = \frac{TN}{TN+FN}$$

CI W/ WOOLF'S METHOD

- RR : $\log(RR) \pm z \cdot \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$
- OR : $\log(OR) \pm z \cdot \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

} then exp LB and UB

CONFOUNDING

When we are talking about the association btw outcome and exposure, we have to take into account the impact of other variables

① is a risk factor
② is outside of the causal chain w/ exposure
③ The association must occur in the absence of exposure

③ POISSON

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

β_0 = log of the intercept
$\exp(\beta_0) = \lambda$ is the mean m of y when all $x = 0$
expected value
$\exp(\beta_1)$ = RATE RATIO of the expected m of event

OFFSET = log of the population, to make comparison of rates
↳ transform counts in rates
QUASI POISSON = if residual deviance is very large compared to 1 means that the model needs additional var, and if we don't have those we use quasi poisson, that generates a new coeff estimated, has a > SE and larger CI

$$\frac{\text{residual dev / new coeff est}}{\text{residual df}}$$

ZERO INFLATED MODEL = zero inflation is a common issue w/ Poisson
It's an excess of zero counts that can alter outcome and results ($\downarrow \lambda$)

the zero inflated model removes the bad effect of those zeros

INTERACTIONS = or effect modification
is an additional variable that can change the relation btw exposure and outcome

$$\log \frac{\pi}{1-\pi} : \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \gamma_{12} x_1 x_2$$

MEASURING GOF → good mess of fit

④ LRT : For nested models
We start from model 1: $y = \beta_0 + \underbrace{\beta_1 x_1 + \ldots + \beta_K x_K}_{K}$
We add to mod 1 some var: $y = \beta_0 + \underbrace{\beta_1 x_1 + \ldots + \beta_K x_K}_{K} + \underbrace{\beta_{K+1} x_{K+1}}_{h} \rightarrow h$

OUR Ho : $\beta_{K+1} = \beta_{K+2} = \ldots = \beta_{K+h} = 0$

$D = -2\{\log L(\text{smaller}) - \log L(\text{larger})\}$ if D > Chi² or pvalue we reject Ho

② RESIDUAL DEV / DEG OF FREED — the ideal value is 1; > 1 means that the model can't explain all

res dev = is the variability that the model can't compute

③ AIC = if we don't have nested models

$$AIC = \underbrace{2\log L}_{\text{GOF}} + \underbrace{2m}_{\text{complexity}}$$

THE SMALLER, THE BEST