Week 4

Tuesday, 3 August 2021 22:23

HANDLING BINARY DATA

QUESTION: YES OR NO?

- DOES A PERSON HAVE A CERTAIN DISEASE?
- DOES THE SUBJECT HAVE A SPECIAL CHARACTERISTIC?
- DOES SUBJECT ENGAGE IN A CERTAIN BEHAVIOUR?

EASIER TO SUMMARIZE DATA

CONTINOUS DATA (SPREAD, CENTER, LOCATION)

BINARY DATA (SAMPLE PROPORTION)

WE JUST HAVE 1 SUMMARY STATISTIC

THERE ARE DIFFERENT WAY TO COMPARE DATA

BINARY DATA (DEFINITION & SUMMARIZATION)

$$\hat{p} = \frac{\#\ \text{POSITIVE OUTCOMES}}{\#\ \text{TOTAL OUTCOMES}}$$

"P-HAT"
PROPORTION

$\hat{p} \neq p$

$\hat{p}$ → SAMPLE ESTIMATE
p → TRUE VALUE

VERY SIMILAR TO A MEAN

$$x = \begin{cases} 0 & \text{NEGATIVE OUTCOME} \\ 1 & \text{POSITIVE OUTCOME} \end{cases}$$

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$s = \sqrt{N\hat{p}(1-\hat{p})}$$

STANDARD DEVIATION

- NOT IMPORTANT FOR UNDERSTANDING THE DISTRIBUTION
- USEFUL WHEN COMPARING 2 DIFFERENT DATA

NORMAL DISTRIBUTION: $s_{(\delta)}$ DOES NOT DEPEND ON $\bar{x}_{(\mu)}$

BINOMIAL DISTRIBUTION: $s_{(\delta)}$ DOES DEPEND ON $\bar{x}_{(\hat{p})}$

PERCENTILES WILL EITHER BE 0 OR 1.

EXAMPLE 1000 PATIENTS, 208 HIV POSITIVE

$\hat{p} = \frac{208}{1000} = 0.208 = 20.8\%$ (NOT VERY USEFUL)

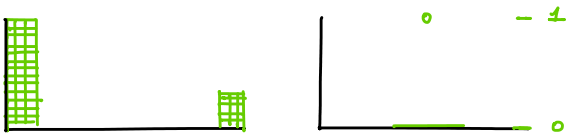
0 0 0 0 0 ... 0 (792) | 1 1 1 1 ... 1 (208)

$q_{0.25} = Q_1 = 0$ $\quad x_{(250)} = 0$
$q_{0.5} = Q_2 = m = 0$ $\quad x_{(500)} = 0$
$q_{0.75} = Q_3 = 0$ $\quad x_{(750)} = 0$

$q_{0.793} = 1$ $\quad x_{(793)} = 1$
$q_{0.95} = 1$ $\quad x_{(950)} = 1$

$\hat{p}$ CAN BE USE TO KNOW BOTH THE
- CENTER
- VARIABILITY
- PERCENTILES

VISUAL DISPLAY ? NOT USEFUL



UNLIKE CONTINOUS DATA :
IF WE HAVE $\hat{p}$ WE HAVE THE ENTIRE STORY

PERCENTAGE,
PROPORTION,
PROBABILITY,
RISK

COMPARING BINARY OUTCOME BETWEEN TWO POPULATIONS

HAVING $\hat{p}_1$, $\hat{p}_2$

DIFFERENCE IN PROPORTIONS $\quad \hat{p}_1 - \hat{p}_2 = D$ $\quad$ RD

RISK DIFFERENCE
ATTRIBUTABLE RISK) → EPIDEMIOLOGY

D % GREATER RESPONCE TO THERAPY IN GROUP 1 COMPARED TO GROUP 2

(ABSOLUTE RISK) EVEN IF IT'S A GOOD THING

RELATIVE RISK
PROPORTIONS RATIO

$$\hat{r} = \frac{\hat{p}_1}{\hat{p}_2}$$

RR

GROUP 1 HAS $\hat{r}$ TIMES MORE RISK THAN GROUP 2

$\hat{r} = 1.52 \longrightarrow$ 52% GREATER RISK

OR $\quad \frac{p_1 - p_2}{p_2}$

RISK DIFFERENCE

CAN BE INTERPRETED AS IMPACT (ASSUMING CAUSATION) ON A FIXED # OF PERSONS

RELATIVE RISK

IMPACT (ASSUMING CAUSATION) AT THE INDIVIDUAL LEVEL

THEY WILL ALWAYS AGREE ON DIRECTION ASSOCIATIONS

RELATIVE RISK TENDS TO BE BIGGER
MEDIA OFTEN USES THEM

ODDS OF AN OUTCOME

$$\hat{\text{ODDS}} = \frac{\hat{p}}{1-\hat{p}}$$

$$\frac{P(\text{EVENT OCCURS})}{P(\text{EVENT DOES NOT OCCUR})}$$

HIGHER ODDS
HIGHER PROPORTIONS

AS $\hat{p} \longrightarrow 1 \quad \hat{\text{ODDS}} \longrightarrow \infty$

ODDS RATIO

$$OR = \frac{\hat{\text{ODDS}}_1}{\hat{\text{ODDS}}_2} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}}$$

LESS INTUITIVE
USED IN CASE/CONTROL STUDIES

CANNOT DIRECTLY ASSESS THE RISK OF INTEREST

NOT A DIRECT COMPARISON RISK BUT A FUNCTION OF THE RISKS CALL ODDS

ALL 3 RISKS AGREE ON DIRECTION
BUT NOT ALWAYS ON MAGNITUDE

$$\hat{RD} = \hat{p}_1 - \hat{p}_2$$

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$$

$$\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)}$$

EXAMPLE $\hat{p}_1 = 0.07$ $\hat{p}_2 = 0.22$

$\hat{RD} = \hat{p}_1 - \hat{p}_2 = -0.15$ (−15%)

INV $\hat{RD} = \hat{p}_2 - \hat{p}_1 = +0.15$ (+15%)

SAME MAGNITUDE CHANGES SIGN

$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2} = 0.32$ (−68%)

INV $\hat{RR} = \frac{\hat{p}_2}{\hat{p}_1} = 3.1$ (+210%)

$\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = 0.27$ (−73% IN ODDS)

INV $\hat{OR} = \frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)} = 3.7$ (+270% IN ODDS)

VERY DIFFERENT

WHY DIFFERENT MAGNITUDE IF DIRECTION IS REVERSED ?

SCALES OF RATIO IS NOT SYMMETRIC

$\ln x$ EQUALIZES THE VALUES (REGARDLESS OF DIRECTION)

$$\ln\left(\frac{x}{y}\right) = \ln x - \ln y$$

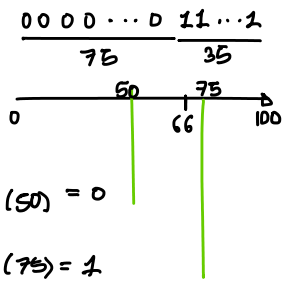
$$\ln(\hat{RR}) = \ln\left(\frac{\hat{p}_1}{\hat{p}_2}\right) = \ln \hat{p}_1 - \ln \hat{p}_2 = k$$

$$\ln(\hat{RR}_{INV}) = \ln\left(\frac{\hat{p}_2}{\hat{p}_1}\right) = \ln \hat{p}_2 - \ln \hat{p}_1 = -k$$

EXERCISES

- $\hat{p} = \bar{x}$ SINCE $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- IF $\hat{p} = 0.35$ 35% YES 65% NO

IF N = 100



$q_{0.50} = m = x_{(50)} = 0$

$q_{0.75} = Q_3 = x_{(75)} = 1$

- $\hat{RR} = 0.78$ $\frac{\text{DRUG}}{\text{PLACEBO}}$

−22% LOWER RELATIVE RISK IF YOU TAKE THE DRUG

- $\hat{RR} = 3$ 3X MORE RISK FOR WHO HAS THAT GENETIC MUTATION
- ABSOLUTE RISK DIFFERENCE $\hat{RD}_{\text{CANCER}} = p_{\text{GENETIC MUTATION}} - p_{\text{NO MUTATION}} = 0.002$

(ASSUMING CAUSALITY) IF N = 10'000
WE WILL HAVE 20 MORE CANCER IF THE WHOLE POPULATION HAS THE MUTATION

- IF $\hat{ODDS} = 1$, $p = ?$

$ODDS(p) = \frac{p}{1-p} = 1 \qquad 1 - p = p$

$2p = 1 \qquad \hat{p} = 0.5 = 50\%$

- STUDY

X BLOOD LEAD LEVELS (MALES) $\sim N(1.97, 1.61^2)$

Y BLOOD LEAD LEVELS (FEMALES) $\sim N(1.27, 1.16^2)$

$N_{X_1} = 100$
$N_{X_2} = 2128$

$S_N \lessgtr S_{100}$? CANNOT KNOW

$\delta = \bar{y} - \bar{x} = -0.7$ mg/dL

"HIGH BLOOD LEAD LEVELS": $X, Y > 3.10$

$B \sim \begin{cases} 0 & BLL < 3.10 \\ 1 & BLL > 3.10 \end{cases}$

$\hat{P}_{F,\text{HIGH BLL}} = 5\%$ $\quad \hat{RR} = \frac{\text{FEMALES}}{\text{MALES}}$ $\quad \underline{\hat{P}_{M,\text{HIGH BLL}}\ ?}$

$\hat{P}_{M,\text{HIGH BLL}} = 3\,\hat{P}_{F,\text{HIGH BLL}} = 15\%$