Correlated Data

4/12/22

Paired Data – Binary Response

STAT 216 → Paired t-test:
- Quantitative (normal) response
- Measured in pairs

→ Pre / Post
→ Husband / Wife

2×2 Table:

Paired

2nd member of pair (Post)

1st member of pair (pre)	Yes	No	
Yes	n_{11}	n_{12}	n_{1+}
No	n_{21}	n_{22}	n_{2+}
	n_{+1}	n_{+2}	n

↳ "Y" on both rows & cols

independent samples

Y = go for walks more than 2×/week

X = own a pet

X	0	1	
0			
1			
			n

(Sec. 8.1)

McNemar's Test of Independence for Correlated Proportions:

H_0: $P(Y_1 = \text{Yes}) = P(Y_2 = \text{Yes})$

H_a: $P(Y_1 = \text{Yes}) \quad P(Y_2 = \text{Yes})$

one of $<$, $>$, $\neq$ depending on research quest.

$\pi_{ij} = P(\text{being in the } i^{th} \text{ row} \,\&\, j^{th} \text{ col})$

$\Rightarrow P(Y_1 = \text{Yes}) = \pi_{11} + \pi_{12} = \pi_{1+}$

$P(Y_2 = \text{Yes}) = \pi_{11} + \pi_{21} = \pi_{+1}$

– Only interested in "discordant" pairs – individuals in $(1,2)^{th}$ cell or $(2,1)^{th}$ cell

We say there is marginal homogeneity if

$$\pi_{12} = \pi_{21}$$

Why? $\pi_{1+} = \pi_{+1} \longrightarrow$

$$\pi_{1+} - \pi_{+1} = (\cancel{\pi_{11}} + \pi_{12}) - (\cancel{\pi_{11}} + \pi_{21}) = \pi_{12} - \pi_{21}$$

Let $n^* = n_{12} + n_{21} \longrightarrow$ total # of discordant pairs

Then Under H_0, $n_{12} \sim Bin(n^*, 0.5)$.
→ Use to get p-value

or $n^* > 10$ (arbitrary rule of thumb),

$$Z = \frac{n_{12} - n^*(0.5)}{\sqrt{n^*(0.5)(1-0.5)}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12}+n_{21}}} \mathrel{\dot\sim} N(0,1)$$

or $Z^2 \mathrel{\dot\sim} \chi^2(1) \longleftarrow R$

Example: (Pediatrics, 2006) – Fire Safety
Researchers compared children's reactions to two kinds of alarms:
1. Conventional smoke alarm
2. Recording of mother's voice saying the child's name & urging them to wake up.

→ each child tested by each alarm → 2 measurements per child

		Mother's Voice: Awoke	Mother's Voice: Did Not Awake	
Conventional Alarm	Awoke	14	0	14
	Did not awake	$\boxed{9}$	1	10
		23	1	24

observed test stat

$H_0: \pi_{12} = \pi_{21}$ $\Big(\underbrace{P(\text{Awake} \mid \text{Alarm } 1)}_{\pi_{1+}} = \underbrace{P(\text{Awake} \mid \text{Alarm } 2)}_{\pi_{+1}} \Big)$

$H_a: \pi_{12} \neq \pi_{21}$

p-value $= P(X = 0 \text{ or } X = 9 \mid X \sim Bin(9, 0.5))$

Let $X = n_{12}$ (prior to data collection)

R: sum(dbinom(c(0,9), 9, 0.5))

$= 0.0039063$

$\Rightarrow$ Strong evidence that the probability of waking differs between the two alarm types.

Approximate p-value:

$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{0 - 9}{\sqrt{9}} = -3$$

R: 2 * pnorm(-3) = 0.0026998

or pchisq((-3)^2, 1, lower.tail = FALSE)

		Mother's Voice: Awoke	Mother's Voice: Did Not Awake	
Conventional Alarm	Awoke	14	$\boxed{0}$	14
	Did not awake	$\boxed{9}$	1	10
		23	1	24

observed test stat

How would the data be organized in a 2x2 table if they were incorrectly treated as independent measurements.

ID	Alarm Type	Awoke or Not
1	Conv.	Yes
1	Mother	Yes
2	⋮	⋮
2		
⋮		

→

		Awoke or Not: Yes	Awoke or Not: No	
Alarm Type	Conv			
	Mother			
				48

Asymptotic CI for $\pi_{1+} - \pi_{+1}$: (dependent data)

$$SE(\hat{\pi}_{1+} - \hat{\pi}_{+1}) = \frac{1}{n}\sqrt{(n_{12}+n_{21}) - \frac{(n_{12}-n_{21})^2}{n}}$$

$$CI: \quad \hat{\pi}_{1+} - \hat{\pi}_{+1} \pm z^* \cdot SE(\hat{\pi}_{1+} - \hat{\pi}_{+1})$$

from $N(0,1)$

Asthma Example → 95% CI (0.01891, 0.03709)

We are 95% confident that the true difference in proportions of 13-year-olds having asthma & 20-year-olds having asthma (13 − 20) is between (0.019, 0.037), among children similar to those in the sample.

We are 95% confident that the risk of asthma at age 13 is between 0.019 & 0.037 higher than the risk of asthma at age 20 ...

Notes on "correct = FALSE" argument in R.

- When using a normal approximation (continuous) on discrete data, we often employ a "continuity correction" –

McNemar : $$Z_c^2 = \frac{(|n_{12}-n_{21}| - 0.5)^2}{n_{12}+n_{21}} \underset{H_0}{\overset{\circ}{\sim}} \chi^2_1$$

Yates

Chi-squared : $$X_c^2 = \sum \frac{(|obs - exp| - 0.5)^2}{exp}$$

More Generally – Modeling Framework

Option 1: Population-averaged models – Marginal models

$$P(Y_1=1) = \alpha + \delta$$
$$P(Y_2=1) = \alpha$$
$$\delta = P(Y_1=1) - P(Y_2=1)$$

$$P(Y_t=1) = \alpha + \delta x_t$$
$$x_t = \begin{cases} 1 & t=1 \\ 0 & t=2 \end{cases}$$

or another common model:

$$\text{logit}(P(Y_t=1)) = \log\left(\frac{P(Y_t=1)}{1-P(Y_t=1)}\right) = \alpha + \delta x_t$$

→ But data are not independent. "Clustered" data.

– Need different method for fitting these models → generalized estimating equations (GEE)

→ "Naive SEs" → treating data as independent

* "Robust SEs" → account for correlation