

Correlated Data

4/12/22

Paired Data - Binary Response

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

→ Pre / Post

→ Husband / Wife

- Measured in pairs

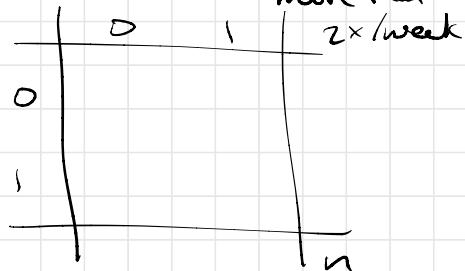
→ 2x2 Table:

Paired

		2 nd member pair (Post)		own per
		Yes	No	
1 st member pair (Pre)	Yes	n_{11}	n_{12}	n_{1+}
	No	n_{21}	n_{22}	n_{2+}
		n_{+1}	n_{+2}	n

independent samples

$\gamma = \text{go for walks more than}$



↳ "Y" on both rows ~ cols

(See 8.1)

McNemar's Test of Independence for Correlated Proportions:

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

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

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

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

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

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

- Only interested in "discordant" pairs
Individuals in $(1, 2)^{\text{th}}$ cell or $(2, 1)^{\text{th}}$ cell

We say there is marginal homogeneity if

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

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

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

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

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

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

$$z = \frac{n_{12} - n^*(0.5)}{\sqrt{n^*(0.5)(1-0.5)}} \stackrel{D}{\sim} \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \sim N(0, 1)$$

or $-z^2 \sim \chi^2(1) \leftarrow R$

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.

\rightarrow each child tested by each alarm \rightarrow 2 measurements per child

		Mother's Voice		
		Awake	Did Not Awake	
Conventional Alarm	Awake	14	0	14
	Did not awake	9	1	10
		23	1	24

$$H_0: \pi_{12} = \pi_{21} \quad (P(\text{Awake} | \text{Alarm 1}) = P(\text{Awake} | \text{Alarm 2}))$$

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

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

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

$$R: \text{sum(dbinom}(c(0, 9), 9, 0.5))$$

$$= \underline{\underline{0.0039063}}$$

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

Approximate p-value:

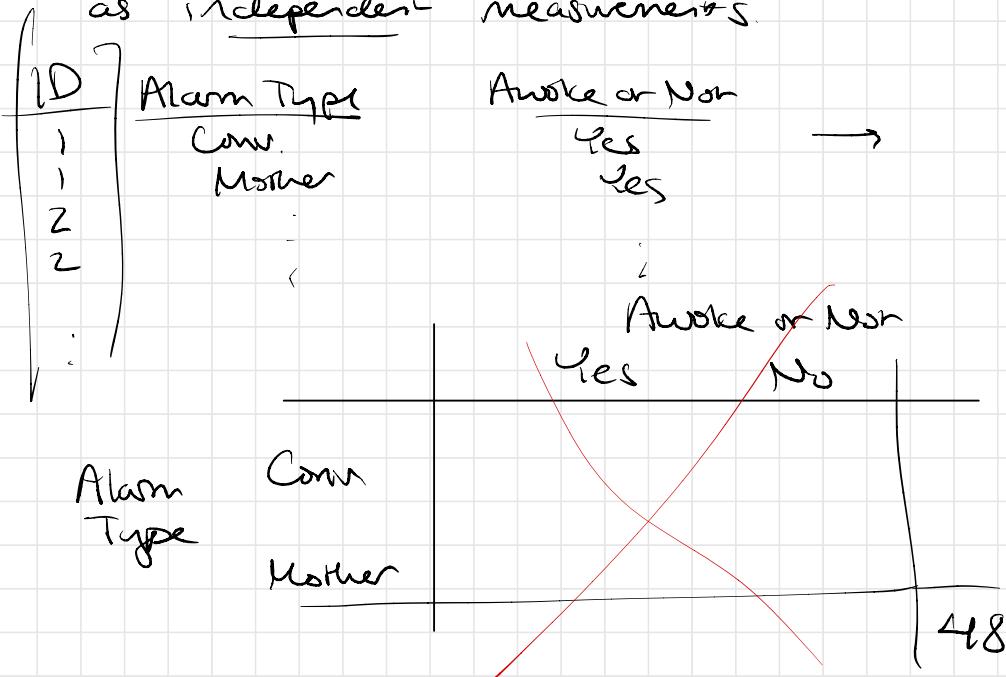
$$Z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{0 - 9}{\sqrt{9}} = -3$$

$$R: 2 * \text{pnorm}(-3) = 0.0026998$$

$$\text{or } \text{pchisq}((-3)^2, 1, \text{lower.tail} = \text{FALSE})$$

		Mother's Voice		
		Awake	Did Not Awake	
Conventional Alarm	Awake	14	0	14
	Did not awake	9	1	10
		23	1	24

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



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

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

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

from $N(0, 1)$

Asthma Example \rightarrow 95% CI $(0.01891, 0.03709)$

We are 95% confident that the difference in proportions of 13-year-olds having asthma vs 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: $Z_c^2 = \frac{(|n_{12} - n_{21}| - 0.5)^2}{n_{12} + n_{21}}$

$\xleftarrow{\text{Takes}} \quad H_0 \quad \sim \chi_1^2$

Chi-squared: $\chi_c^2 = \sum \frac{(\text{obs} - \text{exp})^2}{\text{exp}}$

More Generally - Modeling Framework

Option 1: Population-averaged models — Marginal models

$$\left. \begin{array}{l} P(Y_1=1) = \alpha + \delta \\ P(Y_2=1) = \alpha \\ S = P(Y_1=1) - P(Y_2=1) \end{array} \right\} \quad \left. \begin{array}{l} P(Y_t=1) = \alpha + \delta x_t \\ X_t = \begin{cases} 1 & t=1 \\ 0 & t=2 \end{cases} \end{array} \right.$$

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

