# 1 Models for Matched Pairs

Matched pairs occur when we analyse samples such that for each measurement in one of the samples there is a measurement in the other sample that directly relates to the measurement in the first sample.
Matched pairs typically occur when for example two measurements are taken for the same individual, e.g. response before and after intervention.
If the measurements are categorical, a natural way to display such data would be a contingency table.

**Example 1**
For instance (this is an example from the online notes by Jason Newsom at Portland State University), we might examine the favorability of voters for gun control legislation in April and in June.

| | | June | | |
|---|---|---|---|---|
| | | Yes | No | Total |
| April | Yes | 110 | 10 | 120 |
| | No | 100 | 80 | 180 |
| | Total | 210 | 90 | 300 |

We now could ask if the opinions have changed from April to June. Then we are not interested to test for independence, because if every participant kept their opinion all measurements would be in the cells on the diagonal.
We also might want to compare the proportion of voters who agree with gun control in April and June. For the distribution this would mean comparing $\pi_{yes+}$ with $\pi_{+yes}$.

**In general:** Let $\pi_{ij}$ be the joint distribution of a $2 \times 2$ table. Then the difference between $\pi_{1+}$ and $\pi_{+1}$ is of interest. It is

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

Therefore in fact this becomes a question about the off diagonal entries. To answer if the marginal distributions are equal we can check if $\pi_{12} = \pi_{21}$.

## 1.1 McNemar Test

To test if $\pi_{12} = \pi_{21}$ only the cell counts for those two cells are needed, and the relevant sample size is $n^* = n_{12} + n_{21}$.
With this data we now test if the probability (to fall in cell 12, given it is either in cell 12 or 21) of a binomial distribution is 0.5. Applying the Z-score for binomial distributions to this setting, we get

$$\begin{aligned} Z &= \frac{n_{12} - n^*(0.5)}{\sqrt{n^*(0.5)(0.5)}} = \frac{n_{12} - 0.5(n_{12} + n_{21})}{0.5\sqrt{n_{12} + n_{21}}} = \frac{0.5n_{12} - 0.5n_{21}}{0.5\sqrt{n_{12} + n_{21}}} \\ &= \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \end{aligned}$$

which is approximately standard normally distributed, if the true probability is 0.5.
The McNemar statistic $\chi^2 = Z^2$ is therefore $\chi^2$-distributed with $df = 1$.

To test if the marginal distributions are the same for paired samples with binary response:
**McNemar Test**

1. $H_0 : \pi_{12} - \pi_{21} = 0$ or $H_0 : \pi_{1+} - \pi_{+1} = 0$ versus $H_0$ is not true. Choose $\alpha$.
2. Random sample for the table, $n_{12} + n_{21} \geq 10$.
3. Test Statistic
$$\chi^2_0 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}, \quad df = 1$$
4. P-value $= P(\chi^2 > \chi^2_0)$

**Example 2**
Test if the proportion of people who support gun control has changed from April to June.

1. $H_0 : \pi_{12} - \pi_{21} = 0$ versus $H_0$ is not true. $\alpha = 0.05$.
2. The data represents a random sample, and $n_{12} + n_{21} = 110 \geq 10$.
3. Test Statistic
$$\chi^2_0 = \frac{(90)^2}{110} = 73.6, \quad df = 1$$
4. P-value $= P(\chi^2 > 73.6) < 0.005$
5. Reject $H_0$.
6. At significance level of 5% the data provide sufficient evidence that the proportion of voters who support gun control has changed.

## 1.2 Estimating the difference between two proportions based on paired samples

In general a confidence interval is preferable over a test.
The confidence interval can be based on the difference in the sample proportions:

$$\hat{\pi}_{i+} - \hat{\pi}_{+i}$$

which is an unbiased estimator for the true difference in the probabilities, s.t.

$$\mu_{\hat{\pi}_{i+} - \hat{\pi}_{+i}} = \pi_{i+} - \pi_{+i}$$

and has estimated standard deviation

$$SE = \sqrt{\frac{\hat{\pi}_{i+}(1 - \hat{\pi}_{i+}) + \hat{\pi}_{+i}(1 - \hat{\pi}_{+i}) - 2(\hat{\pi}_{11}\hat{\pi}_{22} - \hat{\pi}_{12}\hat{\pi}_{21})}{n}}$$

Rewriting the $SE$ in terms of observed cell counts gives

$$SE = \frac{\sqrt{(n_{12} + n_{21}) - 2(n_{12} - n_{21})^2/n}}{n}$$

**McNemar Confidence Interval for $\pi_{i+} - \pi_{+i}$**

$$(\hat{\pi}_{i+} - \hat{\pi}_{+i}) \pm z_{\alpha/2} \frac{\sqrt{(n_{12} + n_{21}) - 2(n_{12} - n_{21})^2/n}}{n}$$

**Example 3**
A 95% comfidence interval for the difference in the proportion of voters supporting gun control:

$$\left(\frac{120}{300} - \frac{210}{300}\right) \pm 1.96 \frac{\sqrt{(10 + 100) - 2(10 - 100)^2/300}}{300}$$

$$(0.4 - 0.7) \pm 1.96 \frac{\sqrt{(110-54)}}{300}$$

$$-0.3 \pm 0.0489$$

$$[-0.3489, -0.2511]$$

We are 95% confident that the percentage of voters who support gun control increased between 25% and 35% (based on a sample where the same people were asked in April and June).

## 1.3 Logistic Regression Models for Matched Pairs

### 1.3.1 Marginal Models

In a first step we will model the situation above.
Let $Y_1, Y_2$ be the random variables representing success/failure for observation 1 and observation 2, respectively, and $x_t$ is a dummy variable which is one for observation 1 ($t = 1$) and zero for observation 2 ($t = 2$).
Now analyse the following model

$$logit(P(Y_t = 1)) = \alpha + \beta x_t$$

Then

$$\text{for} \;\; t = 1: \;\; logit(P(Y_1 = 1)) = \alpha + \beta$$

$$\text{for} \;\; t = 2: \;\; logit(P(Y_2 = 1)) = \alpha$$

Which means that $e^\beta$ is the odds ratio for success in observation 1 versus observation 2.

**Example 4**
From the data

$$\Theta = \frac{120/180}{210/90} = 0.2856$$

The odds for support of gun control in April are 0.2856 times the odds in June, or the odds for support of gun control in June are 1/0.2856=3.49 times the odds in April.
Fitting the model gives (inputting the data as if it would be independent, therefore the tests can not be used)

logit(P($Y_t$=Yes)) =$-0.405 - 1.253x_t$ with $e^{-1.253} = 0.2856$.

The model

$$logit(P(Y_t = 1)) = \alpha + \beta x_t$$

is called the *marginal model* because it models the marginal distribution of the response for the paired observations.

### 1.3.2 Conditional models

These models are based on the perception that matched pairs data can be viewed as three way contingency tables, one table for each individual in the sample.

**Example 5**
Assume the first individual in the sample was against gun control in April, but for it in June, then the table for this individuals response would be:

|  | Response |  |
|---|---|---|
| Time | Yes | No |
| April | 0 | 1 |
| June | 1 | 0 |

How many of these tables would we need?

The $k$th table shows the responses $Y_1, Y_2$ for individual $k$, and the sample data is represented as a $2 \times 2 \times n$ table.
Models based on this perception of matched pairs are called *conditional models*, the tables are conditioned on the individual, and permit that the probability distribution for the response is different for each individual.

Notation: Let $Y_{it}$ be observation $t$ for individual $i$ ($1 =$ success).

The conditional model:

$$logit(P(Y_{i1} = 1)) = \alpha_i + \beta, \quad logit(P(Y_{i2} = 1)) = \alpha_i,$$

or

$$logit(P(Y_{it} = 1)) = \alpha_i + \beta x_{it}$$

with $x_{i1} = 1, x_{i2} = 0$.
The model implies a homogeneous relationship because the odds ratio for success for comparing the two observations is $e^\beta$ for all individuals. In the case $\beta = 0$, for each individual the odds (probabilities) for success are the same for the two observations.
The odds (probabilities) for success can be different for individuals depending on $\alpha_i$.

To answer if the probabilities for success are different is equivalent to test within the conditional model if $\beta = 0$.

The many parameters ($n$+1) in the model can cause difficulties with the estimation. One solution is to consider *conditional maximum likelihoods*, where the conditional likelihood function is maximized, by *conditioning out* the individual parameters. For this simple model this leads to the conditional maximum likelihood estimator of $exp(\hat{\beta}) = n_{12}/n_{21}$ from the original 2-way table.

**Example 6**
The conditional maximum likelihood estimate for the odds ratio comparing the support for gun control in April and June equals 10/100=0.10. The odds for support of gun control is in April 10% of the odds in June.

This is very different from the 0.2856 we found using the marginal model, reflecting the difference between marginal and conditional odds ratios.