Violation of independence assumption can cause over-dispersion, e.g., response has a common cause, say a disease is influenced by genes, the responses will tend to be positively correlated.

- Overdispersion cannot arise when \( n=1 \) (sparse case).

\[
\begin{align*}
E(y) &= \sum_i E(s_i) = \sum_i \{ E[E(s_i|\pi_i)] \} = l \times mp = np \\
Var(y) &= \sum_i Var(s_i) = \sum_i \{ E[Var(s_i|\pi_i)] + Var[E(s_i|\pi_i)] \} \\
&= \sum_i \{ E[m\pi_i(1-\pi_i)] + Var(m\pi_i) \} \\
&= l \times \{ mp - m[\tau^2 p(1-p) + p^2] + m^2 \tau^2 p(1-p) \} \\
&= (1+(m-1)\tau^2) np (1-p) \geq np(1-p) \leq \sigma^2 \leq m \leq n
\end{align*}
\]

\[
Var(y) = \sum_i Var(s_i) = \sum_i \{ E[Var(s_i|\pi_i)] + Var[E(s_i|\pi_i)] \} = l \times \{ mp - m[\tau^2 p(1-p) + p^2] + m^2 \tau^2 p(1-p) \} = (1+(m-1)\tau^2) np (1-p) \geq np(1-p)
\]

\[
\begin{align*}
Var(y) &= \sum_i Var(s_i) = \sum_i \{ E[Var(s_i|\pi_i)] + Var[E(s_i|\pi_i)] \} \\
&= \sum_i \{ E[m\pi_i(1-\pi_i)] + Var(m\pi_i) \} \\
&= l \times \{ mp - m[\tau^2 p(1-p) + p^2] + m^2 \tau^2 p(1-p) \} = (1+(m-1)\tau^2) np (1-p) \geq np(1-p)
\end{align*}
\]

- Under-dispersion, e.g., when food supply is limited, survival probability of an animal may be increased by the death of others, i.e., negatively correlated.

Q: how to model overdispersion and do analysis?

- Introduce one additional dispersion parameter \( \sigma^2 \), i.e.,

\[
Var(y_x) = \sigma^2 \times n_x p_x (1 - p_x) \approx \text{notice its similarity to linear model (standard binomial case } \Rightarrow \sigma^2 = 1; \text{ over-dispersion } \Rightarrow \sigma^2 > 1)\]

For a model \( S \), \( \sigma^2 \) can be estimated using \( \hat{\sigma}^2_S \), i.e.,

\[
\hat{\sigma}^2_S = \frac{\text{RSS in } L}{MSE}\]

When \( n_i \)'s large, using \( \hat{\sigma}^2_S \) is OK.

\[
\begin{align*}
\hat{\beta} &\approx \beta \quad (\text{from IRWLS under binomial estimate}) \\
\hat{\sigma}^2 \cdot n_x &\approx \text{quasi-likelihood}
\end{align*}
\]

But, \( Var(\hat{\beta}) \approx \sigma^2 (X^TWX)^{-1} \) and \( Var(\hat{\beta}) = \hat{\sigma}^2 (X^TWX)^{-1} \)

For \( S \) nested in \( L \), difference in their deviances

\[
D_S - D_L \approx \sigma^2 \chi^2_{df_S - df_L} \quad \text{(under } S)\]

When comparing models, e.g., testing \( H_0: S \) vs. \( H_1: L \setminus S \), can use

\[
F = \frac{(D_S - D_L)/(df_S - df_L) \sigma^2_L}{\chi^2_{df_S - df_L}} \approx F_{df_S - df_L, df_L} \quad \text{(under } S)\]

No goodness-of-fit test is possible.

This dispersion parameter method is more appropriate when the covariate classes are roughly equal in size (i.e., \( n_1 \approx n_2 \approx \ldots \approx n_k \)).
• **Matched case-control studies (MCCD):** match each case with one or more controls that have the same or similar values of some potential confounding variables. A group of a case and its corresponding controls is called a *matched set*, e.g.,

<table>
<thead>
<tr>
<th>Case</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st case: age=20; sex=male</td>
<td>2nd case: age=20; sex=female</td>
</tr>
<tr>
<td>n&lt;sup&gt;th&lt;/sup&gt; case: age=70; sex=female</td>
<td></td>
</tr>
</tbody>
</table>

1:M MCCD: *M* controls for each case
- *M* typically small, can vary in size in every matched set
- Each additional control yields a diminished return in terms of increased efficiency in estimating the effects of risk factors
- It is usually not worth exceeding *M*=5

- Some disadvantages of MCCD
  - Lose the possibility of discovering the effects of the confounding variable *W*
  - The data will likely be far from a random sample of the population of interest
• Modeling and Analysis of a 1:M MCCD with \( n \) matched sets
  
  For individual \( i \) in the \( j \)th matched set \( W = w_j \); \( j = 1, ..., n \), observe the value of risk factors \( x_{ij} \)

  Denote \( i = 0 \Rightarrow \text{case} \) and \( i = 1, ..., M \Rightarrow \text{control} \)

  Assume the main-effect model of \( X \) and \( W \) (i.e., no interactions):

  \[
  \logit[p(w_j, x_{ij})] = \logit(p_{ij}) = \eta_{ij} = \alpha_j + x_{ij}^T \beta.
  \]

  Let \( S_j = z_{0j} + z_{1j} + ... + z_{Mj} \) be the number of 1's (i.e., \( D \)) in the \( M+1 \) binary responses observed in the \( j \)th matched set

  \[
  S_j \propto \exp[\sum_j \alpha_j (z_{ij})] \times \exp[\sum_j z_{ij} x_{ij}^T \beta].
  \]

  Regard \( \alpha_j \)'s as nuisance parameters, it is natural to make inference conditional on its sufficient statistics \( S_j \)'s

  Conditional probability of the observed data in \( j \)th matched set

  \[
  P(z_{0j} = 1, z_{1j} = 0, ..., z_{Mj} = 0 \mid S_j = 1) = P(z_{0j} = 1, z_{1j} = 0, ..., z_{Mj} = 0)/P(S_j = 1)
  \]

  \[
  = \frac{\exp(x_{ij}^T \beta)}{\sum_{i=0}^{M} \exp(x_{ij}^T \beta) + \sum_{i=1}^{M} \exp(x_{ij}^T \beta)}
  \]
The complete conditional likelihood is given by

\[ \mathcal{L}(\beta) = \prod_{j=1}^{n} \left\{ 1 + \sum_{i=1}^{M} \exp \left[ (x_{i,j} - x_{0,j})^T \beta \right] \right\}^{-1} \]

- Data from different matched sets are assumed independent.
- Standard likelihood methods can now be employed to make inference.
- The conditional likelihood takes the same form as that used for the proportional hazards model (PHM) in survival/reliability analysis.
- Note that in this approach:
  - \( \alpha_j \)'s are not estimated \( \Rightarrow \) cannot make predictions of \( p_{ij} \)'s.
  - Even if \( \alpha_j \)'s are estimated, they are not likely the estimates of the true effects of \( W \), prediction of \( p_{ij} \)'s is still questionable.
- Only can make statements about the odds ratio as measured by the \( \beta \).