# Quiz

- What are the study designs for case-control study and cohort study?
- State the invariance property.
- If we care about RR (of disease), what study should we use?

# Over Dispersion

- Definition: the variance of the response $Y$ exceeds the nominal variance. $Y \sim Bin(m, \pi) \rightarrow \mathrm{var}_n(Y) = m\pi(1-\pi)$.
- Dispersion parameter:

$$a(\phi) = \mathrm{var}(Y)/\mathrm{var}_n(Y).$$

- Example: we hypothesize that the support rate of NE Patriots is constant across 5 midwestern states. That is, the proportion of people in the populations of those states who would root for the team is constant. We perform polls by randomly sampling $m = 200$ people in each of the 5 states: Wisconsin 57, Michigan 113, Illinois 56, Iowa 121, Minnesota 153.

$Y_i \sim Bin(200, \pi)$ $i = 1, \dots, 5$ $\quad \mathrm{var}_n(Y_i) = 50, \quad \mathrm{var}(Y_i) = 1801$

$\hat{\pi} = \frac{\sum Y_i}{200 \times 5} = 0.5$

$\mathrm{var}_n(Y_i) = 200 \cdot 0.5 \cdot 0.5 = 50$

$\mathrm{var}(Y_i) = \frac{1}{5}\sum_{i=1}^{5}(Y_i - \bar{Y})^2 = 1801$

- Over dispersion is very common in practice (under-dispersion is relatively rare)
- It means the data may not exactly follow the hypothetical distribution
- Grouped data may have over dispersion
- Binary data ($m = 1$) do NOT have over dispersion
- Count data may also have over dispersion

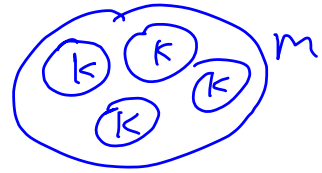
# Source of Over-Dispersion

- Intra-class correlation:
  For $Y = \sum_{i=1}^{m} Y_i$, $Y_i \sim Bin(1, \pi)$, we have $\text{corr}(Y_i, Y_j) = \rho$. Then $Y$ may NOT follow $Bin(m, \pi)$.

$$\begin{aligned} \mathbb{E}(Y) &= m\pi \\ \text{var}(Y) &= m\pi(1-\pi)[1+(m-1)\rho] \end{aligned}$$

if $m > 1$ & $\rho > 0$, $1 + (m-1)\rho > 1$

- Hierarchical sampling:




  Assume $m$ individuals form $m/k$ clusters, each with cluster size $k$. Each cluster has its own event rate which follows a distribution. For example, assume $Y_i|p_i \sim Bin(k, p_i)$ and $p_i \sim (\pi, \delta\pi(1-\pi))$ for $i = 1, \cdots, m/k$. Then $Y = \sum_{i=1}^{m/k} Y_i$ actually has

$$\begin{aligned} \mathbb{E}(Y) &= m\pi \\ \mathrm{var}(Y) &= m\pi(1-\pi)[1+(k-1)\delta] \end{aligned}$$

$\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y|p_i)) + \mathrm{Var}(\mathbb{E}(Y|p_i))$

- In particular, if $k = m$ and $p_i \sim Beta(a, b)$, we have $Y \sim Beta - Binomial(m, \pi, \delta)$, where

$$\begin{aligned} \mathbb{E}(Y) &= m\pi \\ \mathrm{var}(Y) &= m\pi(1-\pi)[1+(m-1)\delta] \end{aligned}$$

  and $\pi = \frac{a}{a+b}$ and $\delta = \frac{1}{1+a+b}$.

# Parameter Estimation

To model over-dispersed binomial data, we assume

$$\mathbb{E}(y_i) = m_i\pi_i, \quad \text{var}(y_i) = m_i\pi_i(1-\pi_i)\phi.$$

We need to estimate $\beta$ and $\phi$.

If $\phi$ is a constant independent of $m_i$

- Quasi-likelihood estimator: $\hat{\beta}_Q = \hat{\beta}_{MLE}$
- $\text{var}(\hat{\beta}_Q) = \phi\text{var}(\hat{\beta}_{MLE})$

# Estimating Constant Dispersion

Assume the dispersion parameter $\phi$ is a constant for independent $m_i$

- Under the dispersion model, the generalized Pearson $\chi^2$ is

$$G = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)\phi} \sim \chi^2(n-p)$$

- $\phi$ can be estimated from

$$\hat{\phi} = G_0/(n-p),$$

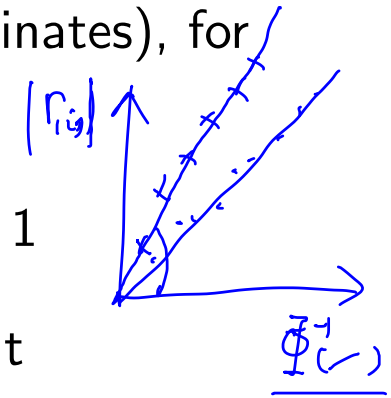
where $G_0$ is the generalized Pearson $\chi^2$ from the original model fitting without over-dispersion.

- Similarly, $\phi$ can be also estimated by $D_0/(n-p)$ where $D_0$ is the deviance of the original model without over-dispersion.

# Model Checking of Over Dispersion

**Half Normal Plot:**

- Order the absolute value of residuals (Pearson or deviance residuals) obtained from the Binomial model without dispersion
- Plot $|r_{(i)}|$ (y coordinates) against $\Phi^{-1}(\frac{n+i+0.5}{2n+1.125})$ (x coordinates), for $i = 1, \cdots, n$
- Reference line is a straight line through origin with slope 1
- Linear deviation from the reference line indicates constant over-dispersion
- Empirical slope is roughly $\sqrt{\phi}$



# Hypothesis Testing

$H_0 : \beta = \beta_0$ vs $H_1 : \beta \neq \beta_0$

- **Wald Test**

$$TS = (\hat{\beta}_Q - \beta_0)^T \text{var}^{-1}(\hat{\beta}_Q)(\hat{\beta}_Q - \beta_0) = TS_W/\phi$$

where $TS_W$ is the Wald Test statistic for the model ignoring dispersion.

- **Score Test**

$$TS = Q(\beta_0)^T \text{var}^{-1}(Q(\beta_0))Q(\beta_0) = TS_S/\phi$$

where $TS_S$ is the Score Test statistic for the model ignoring dispersion.

Assume $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ is the coefficient vector.

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0} \quad \text{vs} \quad H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

**Deviance Analysis**

- Model 1: $\eta = \boldsymbol{Z}_1\boldsymbol{\beta}_1$, Deviance $D_1$
- Model 2: $\eta = \boldsymbol{Z}_1\boldsymbol{\beta}_1 + \boldsymbol{Z}_2\boldsymbol{\beta}_2$, Deviance $D_2$
- Pearson $\chi^2$ statistic $G_0$
- Calculate $D_1, D_2, G_0$ from binomial model without dispersion.
- $\frac{D_1 - D_2}{\phi} \sim \chi^2(p_2)$ and $\frac{G_0}{\phi} \sim \chi^2(n - p_1 - p_2)$ are approximately independent
- Therefore, the test statistic is

$$\frac{(D_1 - D_2)/p_2}{G_0/(n - p1 - p_2)} \sim F(p_2, n - p_1 - p_2) \quad \text{(under } H_0\text{)}$$

# Example

The table below listed the number of rats surviving the 21-day lactation period (Y), and the number of rats alive at four days in the same litter (m) in control and treated groups.

| Group | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| Control | Y | 13 | 12 | 9 | 9 | 8 | 8 | 12 | 11 |
| (X=0) | m | 13 | 12 | 9 | 9 | 8 | 8 | 13 | 12 |
| Treated | Y | 12 | 11 | 10 | 9 | 10 | 9 | 9 | 8 |
| (X=1) | m | 12 | 11 | 10 | 9 | 11 | 10 | 10 | 9 |

| Group | | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|
| Control | Y | 9 | 9 | 8 | 11 | 4 | 5 | 7 | 7 |
| (X=0) | m | 10 | 10 | 9 | 13 | 5 | 7 | 10 | 10 |
| Treated | Y | 8 | 4 | 7 | 4 | 5 | 3 | 3 | 0 |
| (X=1) | m | 9 | 5 | 9 | 7 | 10 | 6 | 10 | 7 |