

Final Review: the Final is on Monday, May 8 12:50-2:50 (here).

The final is cumulative but there is more emphasis on the material in Exam 3 and on quizzes 9 and 10 than on earlier material. 8 sheets of notes.

Material since Exam 3.

Below are the population and observed 2×2 tables.

	Y = 1 = S	Y = 2 = F		Y = 1 = S	Y = 2 = F
X = 1	π_{11}	π_{12}	X = 1	n_{11}	n_{12}
X = 2	π_{21}	π_{22}	X = 2	n_{21}	n_{22}

Let $\pi_1 = \pi_{11} = P(Y = S|X = 1)$ and let $\pi_2 = \pi_{21} = P(Y = S|X = 2)$.

Then in row 1 the odds of a success is $\Omega_1 = \pi_1/(1 - \pi_1) = \pi_{11}/\pi_{12}$,

and in row 2 the odds of a success is $\Omega_2 = \pi_2/(1 - \pi_2) = \pi_{21}/\pi_{22}$.

If the odds

$$\Omega = \frac{\pi}{1 - \pi}, \text{ then } \pi = \frac{\Omega}{\Omega + 1}.$$

The odds ratio is

$$\theta = \frac{\Omega_1}{\Omega_2}.$$

The relative risk equals

$$\frac{P(Y = 1|X = 1)}{P(Y = 1|X = 2)} = \frac{\pi_1}{\pi_2} = \frac{\pi_{11}}{\pi_{21}}.$$

63) The estimated odds ratio is

$$\hat{\theta} = \frac{\hat{\Omega}_1}{\hat{\Omega}_2} = \frac{n_{11}n_{22}}{n_{21}n_{12}}.$$

64) **Unless you are told that the 2×2 table comes from a case-control study,** then the estimated relative risk is

$$\frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{\hat{\pi}_{11}}{\hat{\pi}_{21}} = \frac{n_{11}/(n_{11} + n_{12})}{n_{21}/(n_{21} + n_{22})}.$$

65) **If the table is from a case-control study,** then you can estimate $P(X = 1|Y = 1)$ and $P(X = 1|Y = 2)$ but you can not estimate π_1 and π_2 . Hence the relative risk can not be estimated directly. However, if $\pi_1 < 0.05$ and $\pi_2 < 0.05$ (which is usually true in case control studies), then the estimated odds ratio is used as the estimated relative risk.

66) A 95% CI for $\log(\theta)$ is $\log(\hat{\theta}) \pm 1.96SE(\log(\hat{\theta})) = (L, U)$ where

$$SE(\log(\hat{\theta})) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

67) A 95% CI for θ is (e^L, e^U) where L and U are given in point 66).

68) odds ratio = relative risk $\left(\frac{1 - \hat{p}_2}{1 - \hat{p}_1}\right)$.

Z		Y = 1 = S	Y = 2 = F	summary statistic
1	X = 1	n_{111}	n_{121}	$\hat{\theta}_{XY(1)}$
1	X = 2	n_{211}	n_{221}	
2	X = 1	n_{112}	n_{122}	$\hat{\theta}_{XY(2)}$
2	X = 2	n_{212}	n_{222}	
⋮	⋮	⋮	⋮	⋮
k	X = 1	n_{11k}	n_{12k}	$\hat{\theta}_{XY(k)}$
k	X = 2	n_{21k}	n_{22k}	

A three way table has variables X , Y and Z . Often Y is the response variable, X is an explanatory variable, and Z is a lurking (or latent or confounding) variable, in that the relationship between X and Y is of interest but Z is thought to affect the X - Y relationship. $2 \times 2 \times k$ tables such as the one shown above are of special interest.

The $2 \times 2 \times k$ table has k partial tables. The big table can be collapsed into a 2×2 X_Y marginal table. The associations between X and Y in the partial tables are called conditional associations

69) Simpson's paradox: the marginal X - Y association can have a different direction than the direction of the conditional X - Y associations (e.g. all $\theta_{XY(i)} > 1$ while $\theta_{XY} < 1$).

70) The conditional odds ratio for the j th partial table is

$$\hat{\theta}_{XY(j)} = \frac{n_{11j}n_{22j}}{n_{12j}n_{21j}}.$$

71) 4 step CMH test for conditional independence

- i) Ho: $\theta_{XY(1)} = \dots = \theta_{XY(k)} = 1$ Ha: not Ho
- ii) CMH test statistic (from output)
- iii) p-value = $P(\chi_1^2 > CMH)$.
- iv) If p-value $< \alpha$, reject Ho, X and Y are not conditionally independent given Z otherwise fail to reject Ho, X and Y are conditionally independent given Z .

72) 4 step Breslow-Day test for homogeneity for a $2 \times 2 \times k$ table.

- i) Ho: $\theta_{XY(1)} = \dots = \theta_{XY(k)}$ Ha: not Ho
- ii) BD test statistic (from output)
- iii) df = $k - 1$ and p-value = $P(\chi_{k-1}^2 > BD)$.
- iv) If p-value $< \alpha$, reject Ho, the X - Y association is not homogeneous given Z otherwise fail to reject Ho, there is homogeneous X - Y association given Z .

Consider loglinear models in X, Y and Z . Then the full model is the saturated model (XYZ) is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

The symbol form of the model lists the highest order terms of the model. For example the saturated model is (XYZ) .

73) Given the symbol form of the model, write the model in terms of μ and the λ 's.

74) Given the model in terms of μ and the λ 's, write the model in symbol form.

Now consider loglinear models in i) X and Y or ii) X, Y and Z or iii) W, X, Y and Z . The full = saturated model has $G^2(F) = 0$ and tends to be good if all of the cell counts are large. The independence model (W, X, Y, Z) is usually to simple to fit well. If a model only contains two factor interactions, then a model containing λ^{XY} means that there is an X - Y association, otherwise X and Y are conditionally independent given the remaining variables. 3 way or higher order interactions are hard to interpret.

75) Given a goodness of fit table, as a rule of thumb choose the simplest model that fails to reject H_0 . This rule of thumb is not very good. Let

$$D = \sum \frac{|n_i - \mu_i|}{2n} = \sum \frac{|\hat{p}_i - \hat{\pi}_i|}{n}$$

where the n_i are the observed cell counts and the μ_i are the expected counts under the model M . If $D = D(M) < 0.03$ then the expected counts from the model fit the observed counts well. Hence a better rule of thumb is choose the simplest model M with $D(M) < 0.03$.

76) The 4 step change in deviance test for a reduced model R versus the saturated = full model F is

i) H_0 the reduced model is good H_a use the full model

ii) $G^2(R|F) = G^2(R)$

iii) $df = \text{number of parameters in full model} - \text{number of parameters in reduced model}$
and $p\text{-value} =$

$$P(\chi_{df}^2 > G^2(R|F)).$$

iv) If $p\text{-value} < \alpha$, reject H_0 and use the full model.

If $p\text{-value} \geq \alpha$, fail to reject H_0 and use the reduced model.

Suppose the CMH test fails to reject H_0 . Then X does not affect Y given Z : $\theta_{XY(i)} = 1$ for $i = 1, \dots, k$. If The CMH test is rejected, perform the BD test. If the BD test fails to reject H_0 then there is X - Y dependence (association) given Z and this dependence is the same in all k partial tables (the $\theta_{XY(i)} \equiv \theta$ for $i = 1, \dots, k$. if the BD test reject H_0 , then there is X - Y dependence given Z , but the dependence depends on the level i of Z .