c. Cohort Study

i. Notation: $O_{jk} \sim \mathcal{P}(Q_{jk})$, independent

ii. Are distributions into rows independent of distribution into columns?

- Equivalent to $E_{jk} = E_{j+}E_{+k}/E_{++}$

iii. Use χ^2 test statistic as before

- $T = \sum_{j,k=0}^{1}(O_{jk} - \hat{E}_{jk})^2/\hat{E}_{jk}$
- Expectation satisfies
  - ▷ $\hat{E}_{j+} = O_{j+}$ $\hat{E}_{+k} = O_{+k}$, (3 equations, 4 unknowns)
  - ▷ $\hat{E}_{00}\hat{E}_{11}/(\hat{E}_{10}\hat{E}_{01}) = \psi_0$
  - ▷ If $\psi_0 = 1$ then $\hat{E}_{jk} = O_{j+}O_{+k}/O_{++}$
  - ▷ Hence statistic has distribution χ^2_1
- Equivalently, $T = (O_{00} - \hat{E}_{00})^2/v$ for some v
  - ▷ $v = (\sum \hat{E}_{jk}^{-1})^{-1}$

$$= \left(\frac{O_{++}}{O_{+0}O_{0+}} + \frac{O_{++}}{O_{+0}O_{1+}} + \frac{O_{++}}{O_{+1}O_{0+}} + \frac{O_{++}}{O_{+1}O_{1+}}\right)^{-1}$$
$$= O_{+1}O_{0+}O_{+0}O_{1+}/{O_{++}}^3$$

iv. v above is same as approximation arising from stratified cohort formulation

- Hence approximate inference is same as if we had conditioned on row totals
- This conditioning is suggested by conditionality principal.
- Normal approx. works poorly unless $\hat{E}_{jk} \geq 5 \forall j, k$. See Figure 5.
- Could have continuity correction described earlier.
  - ▷ Choice of cc and variance give 4 possible tests

v. Likelihood ratio

- Write down probability for table as function of ψ
- Compare value at 1 to highest value it takes
- $2 \times \log(L) \sim \chi^2_1$

B&D1: 4.2

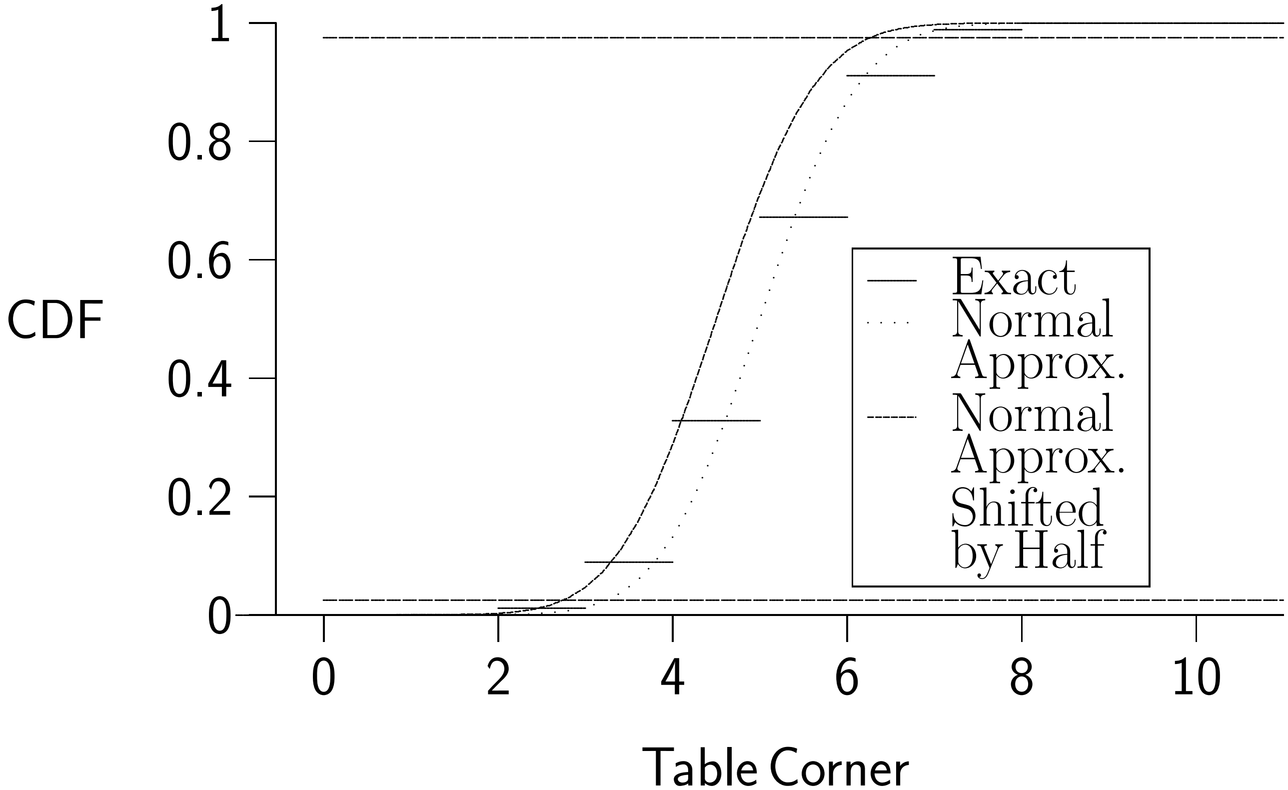
2. Exact Inference for Various Designs

a. As with approximate analysis,

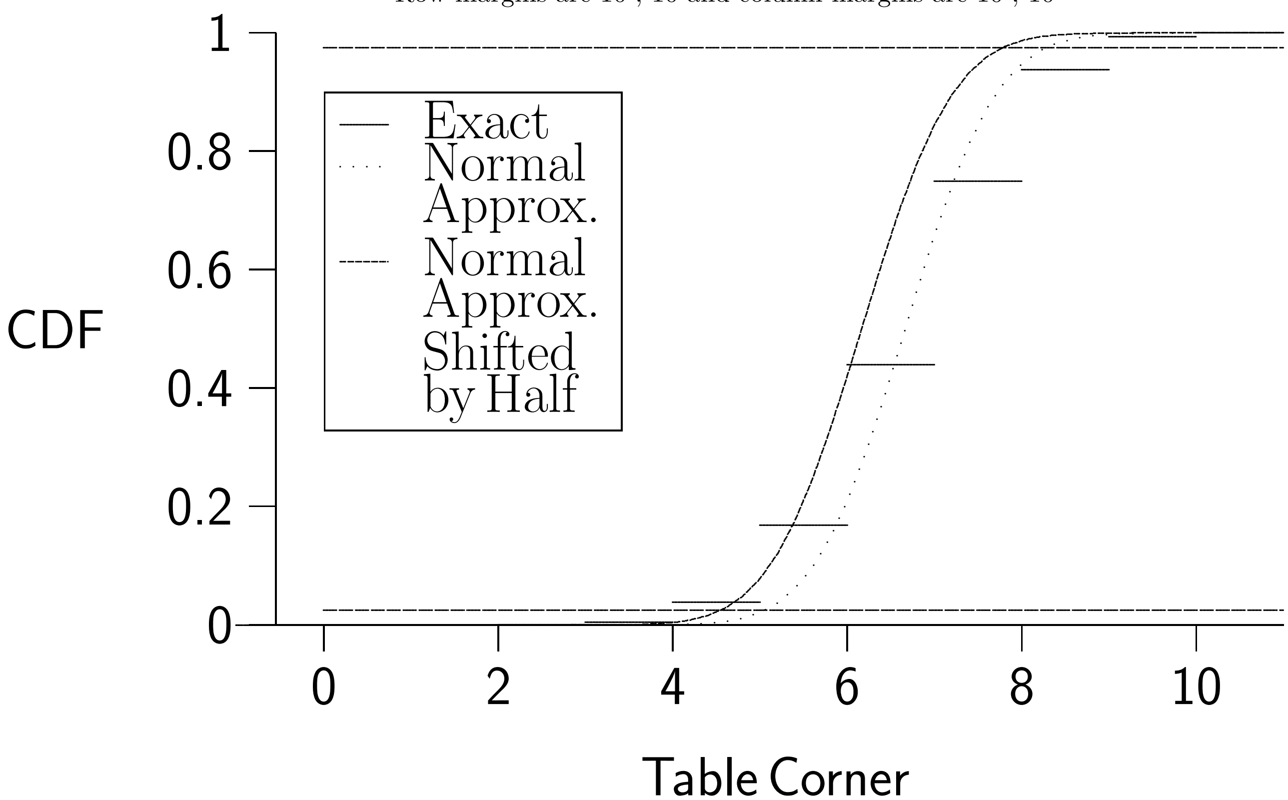
i. case–control approach is mathematically equivalent to the stratified cohort approach

ii. conditionality principal justifies treating the unstratified cohort design as a stratified cohort design.

b. Cohort inference is generated from distribution of



Row margins are 10 , 10 and column margins are 10 , 10



Row margins are 20 , 10 and column margins are 10 , 20

$$O_{00} \sim \mathsf{Bin}(\pi_0, O_{0+})\,,\; O_{10} \sim \mathsf{Bin}(\pi_1, O_{1+})\,.$$

i. π_0 is proportion of exposed PYAR among controls $=$ P [Control|Unexposed]

ii. π_1 is proportion of unexposed PYAR among controls $=$ P [Control|Exposed]

c. $\mathsf{P}\left[O_{00}, O_{10} | O_{0+}, O_{1+}\right] = \binom{O_{0+}}{O_{00}}\binom{O_{1+}}{O_{10}}\pi_0^{O_{00}}(1 - \pi_0)^{O_{01}}\pi_1^{O_{10}}(1-\pi_1)^{O_{11}}$

d. Rewriting in terms of ψ leaves dependence on one of these:

$\pi_1 = \pi_0\psi/(1-\pi_0+\pi_0\psi)$ and

$$\begin{aligned}
\mathsf{P}\left[O_{00}, O_{10} | O_{+0}, O_{+1}\right] &= \binom{O_{0+}}{O_{00}}\binom{O_{1+}}{O_{10}}(1-\pi_1)^{O_{1+}} \\
&\quad \times \pi_0^{O_{+0}}(1-\pi_0)^{O_{01}-O_{10}}\psi^{O_{1+}} \\
&= \binom{O_{0+}}{O_{00}}\binom{O_{1+}}{O_{10}}\left(\frac{1-\pi_0}{1-\pi_0+\pi_0\psi}\right)^{O_{1+}} \\
&\quad \times \pi_0^{O_{+0}}(1-\pi_0)^{O_{+1}-O_{1+}}\psi^{O_{10}} \\
&= \binom{O_{0+}}{O_{00}}\binom{O_{1+}}{O_{10}}\frac{\pi_0^{O_{+0}}(1-\pi_0)^{O_{+1}}\psi^{O_{10}}}{(1-\pi_0+\pi_0\psi)^{O_{1+}}}
\end{aligned}$$

e. Distribution of T still depends on π_0

i. π_0 contributes a constant factor to all tables with same O_{+0}, O_{+1}

ii. Looking only at such tables

iii. Process is conditional on O_{0+} and O_{1+} as well as O_{+0} and O_{+1}.

iv. removes dependence on π

v. Distribution is called $hypergeometric$

vi. If $\psi \neq 1$ called $noncentral\, hypergeometric$

f. cuts number of tables to be examined.

i. Both a blessing and a curse.

- Indicate by $|O_{j+}, O_{+k}$ conditional on O_{0+} and O_{1+} and O_{+0} and O_{+1}.

ii. $\mathrm{Var}_{\psi=1}\left[O_{00}|O_{j+}, O_{+k}\right] = \frac{O_{+1}O_{0+}O_{+0}O_{1+}}{{O_{++}}^2(O_{++}-1)}$

iii. Conditioning is not suggested by conditionality principal.

- P $[\text{disease}] = \pi_0(O_{0+} + O_{1+}\psi/(1-\pi_0+\pi_0\psi))$
- Dependence is weak.

g. Testing ψ

i. One-sided

- $H_0 : \psi = \psi_0$ vs $H_A : \psi > \psi_0$
- Use $T = \hat{\psi}$ or equivalently O_{00}
- p-value is sum of probabilities for table with upper left corner

at least observed

ii. For two–sided test

- order tables according to null probability
- Implies something other than doubling smaller 1-sided p-value
- Result is called *Fisher's Exact Test*

Se: 7 pp. 201–205

h. Confidence Bounds for ψ

i. Distribution of $\hat{\psi}$?

- $\hat{\psi} \approx \mathcal{N}(\psi, ?)$
- For stratified cohort study?
  - ▷ $\log(\hat{\pi}_0/[1-\hat{\pi}_0]) = \log(O_{01}) - \log(O_{00})$
  - ▷ Under unknown ψ, stratified cohort sampling, $\frac{d}{dO_{00}} \log(\textsf{odds}) = {O_{00}}^{-1} + {O_{01}}^{-1}$
  - ▷ $\text{Var}\left[\log(\textsf{odds})\right] \approx ({O_{00}}^{-1} + {O_{01}}^{-1})^2({O_{00}}^{-1} + {O_{01}}^{-1})^{-1} = ({O_{00}}^{-1} + {O_{01}}^{-1})$
  - ▷ Bottom row is independent with same structure
  - ▷ $\text{Var}\left[\hat{\psi}\right] \approx {O_{00}}^{-1} + {O_{10}}^{-1} + {O_{01}}^{-1} + {O_{11}}^{-1}$
- Conditioning on all marginals?

- ▷ No closed form expression for variance
- ▷ Hence $\mathrm{Var}\left[\hat{\psi}\right] \approx {O_{00}}^{-1} + {O_{10}}^{-1} + {O_{01}}^{-1} + {O_{11}}^{-1}$

ii. Hence CI for $\log(\psi)$ is $\log(\hat{\psi}) \pm 1.96 \times \sqrt{{O_{00}}^{-1} + {O_{10}}^{-1} + {O_{01}}^{-1} + {O_{11}}^{-1}}$

iii. Exact Confidence intervals (ψ_L, ψ_U) satisfies

$\mathsf{P}_{\psi_L}\left[O_{00} \geq \text{observed} | O_{j+}, O_{+k}\right] = .025$,

$\mathsf{P}_{\psi_U}\left[O_{00} \leq \text{observed} | O_{j+}, O_{+k}\right] = .025$

- See Figure 4/.

Se: 7 pp. 234–239

3. Controling for the presence of additional variables

a. Notation:

i. Add superscript i to tell which table

Se: 6 pp. 187–190

b. Additional variable provides an alternative explanation for association between disease and exposure: *confounding*

i. Definition: distortion of disease/exposure association by other factor

- Other factor related to exposure

- Other factor causally related to disease

$$\begin{array}{l} C \to D \\ \downarrow \\ E \end{array}$$

ii. Can change direction of relationship: *Simpson's Paradox* (See example)

iii. Rational
- Define the effect of exposure to be that with everything else held constant
- what you will get if you try to intervene on exposure
- This is what you get if you assign exposure

iv. Example
- Aspirin is associated with stomach upset
- Does aspirin cause stomach upset?
- Alternative explanation: stress causes
  - ▷ stomach upset
  - ▷ diseases like headaches for which aspirin is likely treatment.
- Regardless of what book says, you can't tell direction of causation from an observational study

c. Testing whether common odds ratio is 1

i. Use $T = \sum_{i=1}^{I} w_i({O_{11}}^i - E_{11}^i)$

- Intuition might suggest $w_i = 1/\sqrt{\mathrm{Var}\left[{O_{00}}^i | O_{j+}, O_{+k}\right]}$
- We will use $w_i = 1$
- Use as standard error sum of exact variances.
  - ▷ Implies assumption that tables are independent.

ii. Called *Mantel–Haenszel test*.

d. Estimation of the common odds ratio

i. *Mantel–Haenszel estimator* $\frac{\sum_{i=1}^{I} {O_{00}}^i {O_{11}}^i / {O_{++}}^i}{\sum_{i=1}^{I} {O_{10}}^i {O_{01}}^i / {O_{++}}^i}$

ii. ∞ only if all bottom products are 0

iii. *logit estimator*

$$\hat{\psi} = \exp\left(\frac{\sum_{i=1}^{I} w_i \log({O_{00}}^i {O_{11}}^i / [{O_{10}}^i {O_{01}}^i])}{\sum_{i=1}^{I} w_i}\right)$$

- $w_i = (\frac{1}{{O_{00}}^i} + \frac{1}{{O_{01}}^i} + \frac{1}{{O_{10}}^i} + \frac{1}{{O_{11}}^i})^{-1}$
- Omit term i if ${O_{jk}}^i = 0$ for some j, k
  - ▷ $w_i = 0$
  - ▷ Corresponding logit will be ∞
  - ▷ Acceptable since $\lim_{x \to 0} x \log(x) = 0$
- This w_i minimizes variance
- SE of $\log(\hat{\psi})$ is $1/\sum_j w_j$

Se: 6 pp. 163–165

4. K exposure groups, for K possibly greater than 2.

a. Table entries

	Contr.	Cases	Total
Exp. cat. 0	O_{00}	O_{01}	O_{0+}
Exp. cat. 1	O_{10}	O_{11}	O_{1+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Exp. cat. $K-1$	O_{K-10}	O_{K-11}	O_{K-1+}
Total	O_{+0}	O_{+1}	O_{++}

b. Estimation of effect

i. Pick one group as baseline

ii. Calculate odds ratio compared to this group as before

iii. Also can calculate CI

- Via normal theory and same SE or exactly

iv. Remember these things are NOT independent

c. Testing

i. Don't:

- Test pairwise
- because of multiple comparisons problems.

ii. Use same statistic as before

- Calculate expected values $E_{jk} = O_{j+}O_{+k}/O_{++}$
- $T = \sum_{j=1}^{2} \sum_{k=0}^{K-1} (O_{jk} - E_{j,k})^2 / E_{jk}$.

- $T \sim \chi^2_{K-1}$ (approximately)
  - ▷ Same requirement of > 5 expected
  - ▷ Exact methods are available

    This time the test statistic won't correspond to one tail

    Now use Pearson statistic.
- DF are same as number of odds ratios one could estimate.

iii. Could also analyze stratified $2 \times K$ tables.

Se: 6 pp. 167–178

d. Could also treat ordered categories

i. Assign each of the categories a score x_k

- By default these are equally spaced
- Alternatively, one can use $Ridit\, scores\; x_k = [\sum_{j<k} O_{+j} + (O_{+k} + 1)/2]/O_{++}$
  - ▷ Gives Mann–Whitney–Wilcoxon test
  - ▷ Test statistic has interpretation as estimated probability that a random individual from one group has a higher score than random individual from the other

ii. Called $Mantel\text{–}Haenszel\, test$.

iii. Calculate $T = \sum_{k=0}^{K-1} x_k(O_{k1} - e_{k1})$

iv. Multiple of correlation betw. row and column scores (0 and 1):

v. Squaring and rescaling makes it $\approx \chi^2_1$

- Rescaling is done using exact variance
- $\mathrm{Var}\left[O_{k1}\right] = O_{k+}O_{+0}O_{+1}(O_{++} - O_{k+})/({O_{++}}^2(O_{++} - 1))$
- $\mathrm{Var}\left[O_{k1} + O_{j1}\right] = (O_{k+} + O_{j+})O_{+0}O_{+1}(O_{++} - O_{k+} - O_{j+})/({O_{++}}^2(O_{++} - 1))$
- Cov $\left[O_{k1}, O_{j1}\right] = (\mathrm{Var}\left[O_{k1} + O_{j1}\right] - \mathrm{Var}\left[O_{k1}\right] - \mathrm{Var}\left[O_{k1}\right])/2 = -O_{k+}O_{j+}[O_{+0}O_{+1}]/({O_{++}}^2(O_{++} - 1))$
- Hence

$$\mathrm{Var}\left[T\right] = \frac{O_{+0}O_{+1}}{O_{++}(O_{++} - 1)} \left\{ \sum_{k=0}^{K-1} x_k^2 \frac{O_{k+}}{O_{++}} - (\sum_{k=0}^{K-1} x_k \frac{O_{k+}}{O_{++}})^2 \right\}$$

- Formally equivalent to test with ordered categories for SMR
- Treating this as standard least–squares regression gives you reasonable SE for test statistic
  - ▷ Regresssing scores on 0 and 1 gives standard two–sample pooled t test
  - ▷ Squaring $\hat{\beta}$/SE gives χ^2_1 statistic

Se: 6 pp. 181–186

e. When do you need to stratify?

i. Heruristically: when stratifier is a confounder

- That is, it is related to both exposure and disease
- Empirically, the odds ratio will change if both row and column proportions differ according to stratifier.

f. If $\psi = 1$ after stratification, disease and exposure are $conditionally\ independent$.

g. If ψ for the various strata are different, there is an interaction between the confounder and exposure.

i. In the next lecture we'll find out how to measure and test it.

h. Checking for confounding via hypothesis test

i. Procedure

- test for association betw. C and D and betw. C and E ,
- adjust if these are significant

ii. Uses significance as a proxy for strength of effect

iii. To make it work at all, typically make very loose criteria for significance

iv. Should not be used for factors that are not confounders

v. Adjust even if effect mitigated by matching.

Se: 9 pp. 277–279, 289–291

5. Extreme case of stratification: Each has two elements
   a. AKA matching
      i. Can either be case–control pairs or exposed–unexposed pairs
      ii. Let n_{il} = number of pairs with case at exposure level i, control at exposure level l
         - Pairs with the same exposure levels for case and control are called *concordant*.
         - Pairs with different exposure levels for case and control are called *discordant*.

Se: 9 pp. 280–282, 287–289

   b. Assumption (exposed–unexposed pairs):
      i. Let π_k^i be the probability of event in exposure group k for pair i
      ii. Assume $\pi_1^i(1-\pi_0^i)/[\pi_0^i(1-\pi_1^i)] = \psi \forall i$
   c. Use Mantel–Haenszel test
      i. For concordant pairs
         - Expected values are exactly observed
         - Variance is zero
         - Hence contribution is zero

ii. For discordant pairs

- Expected is all $\frac{1}{2}$
- Obsd-expected is
  - ▷ $(1 - \frac{1}{2}) = \frac{1}{2}$ for pairs with $+$ association
  - ▷ $(0 - \frac{1}{2}) = -\frac{1}{2}$ for pairs with - association
- Null variance contribution for pair is
  - ▷ approximately $((\frac{1}{2})^{-1} + (\frac{1}{2})^{-1} + (\frac{1}{2})^{-1} + (\frac{1}{2})^{-1})^{-1} = \frac{1}{8}$
  - ▷ More precisely $\frac{1}{8} \times (2/1) = \frac{1}{4}$

iii. Test statistic is same as test that binomial proportion equals $\frac{1}{2}$

- take $\frac{1}{2}(n_{10} - n_{01})$
- multiply by $\sqrt{4/(n_{10} + n_{01})} = 2/\sqrt{n_{10} + n_{01}}$
- Compare to standard normal

d. Called McNemar's Test